A Conceptual Framework for Data Quality in Knowledge Discovery Tasks (FDQ-KDT): A Proposal

David Camilo Corrales[1, 2*], Agapito Ledezma[2], Juan Carlos Corrales[1]

[1] Telematics Engineering Group, Universidad del Cauca, Campus Tulcán, Popayán, Colombia.
[2] Department of Computer Science and Engineering, Universidad Carlos III de Madrid, Avenida de la Universidad 30, 28911, Leganés, Spain.

* Corresponding author. Tel.: (+57-2)820-9900 ext. 2129; email: dcorrales@unicauca.edu.co
Manuscript submitted February 6, 2015; accepted July 25, 2015.
doi: 10.17706/jcp.10.6.396-405

Abstract: Large Volume of Data is growing because the organizations are continuously capturing the collective amount of data for better decision-making process. The most fundamental challenge is to explore the large volumes of data and extract useful knowledge for future actions through data mining and data science methodologies. Nevertheless these not tackle the issues in data quality clearly, leaving out relevant activities. We proposed a conceptual framework for data quality in knowledge discovery tasks based on CRISP-DM, SEMMA and Data Science, considering the issues of ESE Taxonomy.

Key words: CRISP-DM, data quality framework, data science, ESE taxonomy, FDQ-KDT, Knowledge discovery, SEMMA.

1. Introduction

Data explosion is an inevitable trend as the world is connected more than ever. Data are generated faster than ever, and to date about 2.5 quintillion bytes of data are created daily. This speed of data generation will continue in the coming years and is expected to increase at an exponential level, according to International Data Corporation (IDC) recent survey [1]. The most fundamental challenge is to explore the large volumes of data and extract useful knowledge for future actions through data mining and data science[2], [3].

For a successful process of discovery knowledge from data mining exist recognized methodologies such as CRISP-DM and SEMMA [4], [5] which describe the data treatment. Similarly, the Data Science area searches the knowledge extraction with different approaches as stochastic modeling, probability models, signal processing, pattern recognition and learning, etc. [6]. Although the data mining methodologies and data sciences defined the steps for data treatment, these not tackle the issues in data quality clearly, leaving out relevant activities [3]. It has been agreed that poor data quality in data mining, machine learning and data science will impact the quality of results of analyses and that it will therefore impact on decisions made on the basis of these results.

Therefore, in this paper we proposed a conceptual framework for data quality in knowledge discovery tasks based on ESE taxonomy [7]. This framework is a result of filtering elements of CRISP-DM, SEMMA and Data Science, and checking their suitability to the nature in data mining and machine learning projects. The paper is organized as follows. In the next section, we briefly present the data quality framework definitions, the methodologies CRISP-DM, SEMMA, and Data Science. In Section 3, FDQ-KDT framework is described. Finally, we conclude our paper and show the future works in Section 4.

2. Background

2.1. Data Quality Frameworks (DQF)

The DQF seek to assess areas where poor quality processes or inefficiencies may reduce the profitability of an organization [8]. At its most basic, a data quality framework is a tool for the assessment of data quality within an organization [9]. The framework can go beyond the individual elements of data quality assessment, becoming integrated within the processes of the organization. Eppler and Wittig [10] add that a framework should not only evaluate, but also provide a scheme to analyze and solve data quality problems by proactive management.

2.2. Cross Industry Standard Process for Data Mining (CRISP-DM)

CRISP-DM is a comprehensive data mining methodology and process model that provides anyone—from novices to data mining experts—with a complete blueprint for conducting a data mining project. CRISP-DM breaks down the life cycle of a data mining project into six phases: business understanding, data understanding, data preparation, modeling, evaluation, and deployment [11]. Fig. 1 shows the phases with its respective generic tasks of a data mining process.

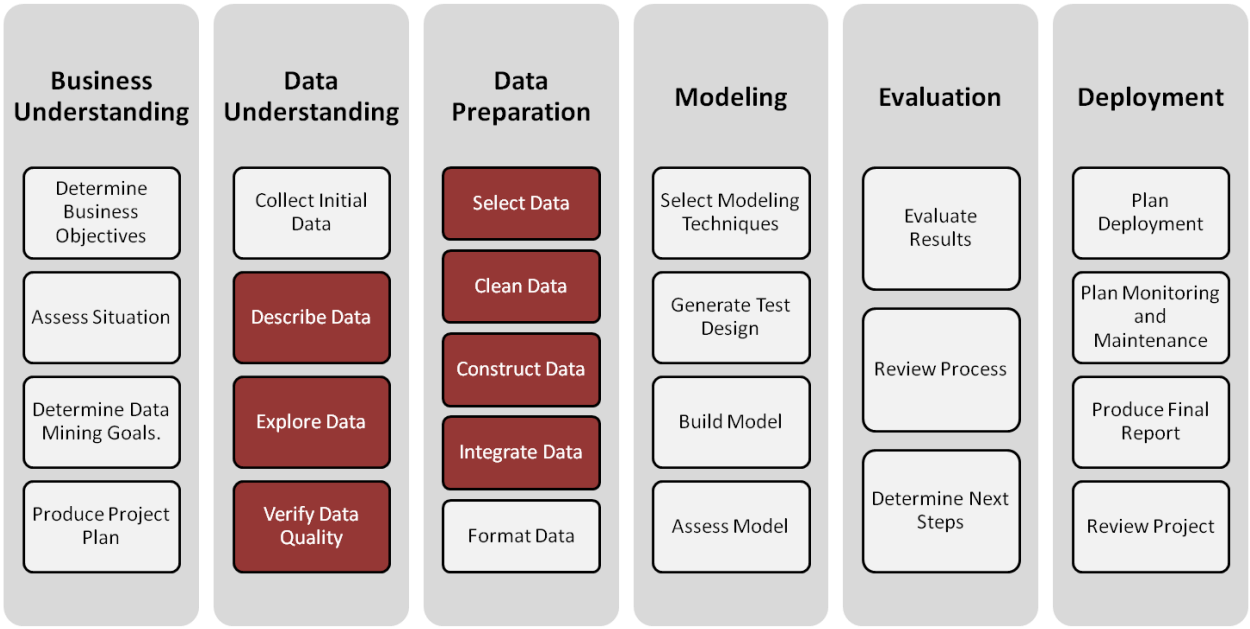

Fig. 1. phases and generic tasks of CRISP-DM.

We considered the following tasks of the phases of *data understanding* and *data preparation* for the construction of FDQ-KDT framework proposed: *describe data*, *explore data*, *verify data quality*, *select data*, *clean data*, *construct data*, and *integrate data*. The remaining tasks of these phases were discarded because we proceed from the fact that FDQ-KDT framework inputs are raw data whereas that the outputs are tidy data in plain text.

2.3. Sample, Explore, Modify, Model and Assess (SEMMA)

The SEMMA process was developed by the SAS Institute that considers a cycle with 5 stages for the process: Sample, Explore, Modify, Model, and Assess. Beginning with a statistically representative sample of your data (sample), SEMMA intends to make it easy to apply exploratory statistical and visualization techniques (explore), select and transform the most significant predictive variables (modify), model the variables to predict outcomes (model), and finally confirm a model's accuracy (assess) [4]. A pictorial representation of SEMMA is given in Fig. 2.

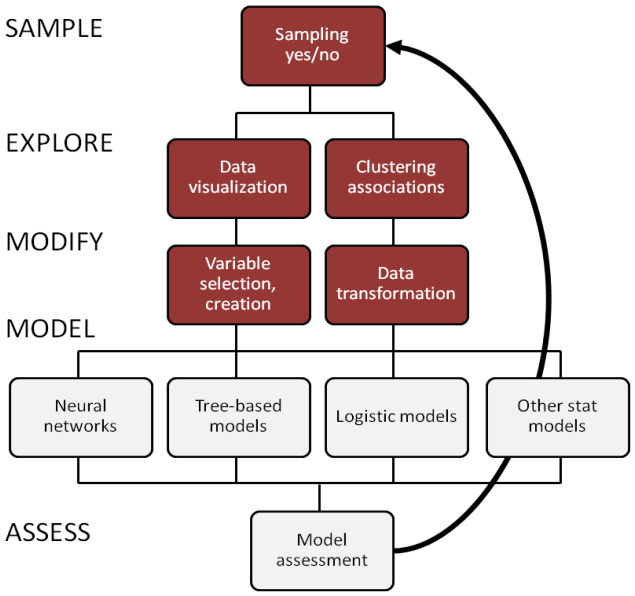

Fig. 2. Stages of SEMMA process.

For FDQ-KDT construction we considered the *sample*, *explore* and *modify* stages of SEMMA process and we discarded the *model* and *assess* stage because the scope of FDQ-KDT is deliver an tidy data in plain text as output.

2.4. Data Science

The data science is focused in the representation, analysis, anomalies of data, and relations among variables [3], from a process with the next steps: raw data collected, data processing, clean data, exploratory data analysis, models and algorithms, construction of reports, and build data product [6]. The data science process flowchart is shown in Fig. 3.

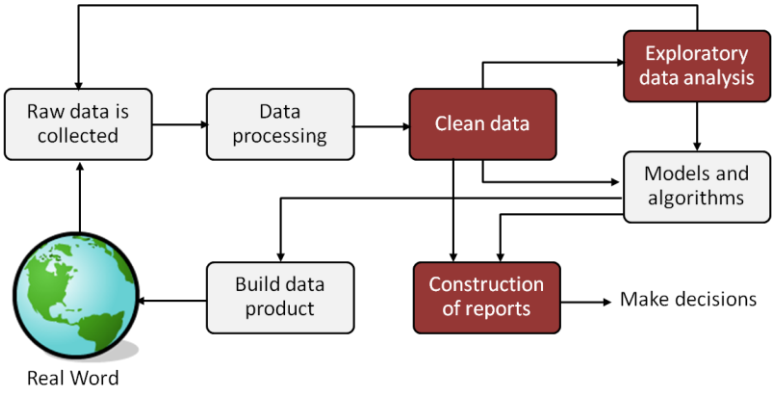

Fig. 3. Data science process flowchart.

We considered the steps: *clean data, exploratory data analysis* and *construction of reports* of data science process flowchart to build the FDQ-KDT. The steps: *raw data is collected*, *data processing, model and algorithms* and *build data product* were not considered because the FDQ-KDT does not aim solve data collection processes and data modeling.

2.5. A Taxonomy of Data Quality Challenges in Empirical Software Engineering (ESE)

The ESE taxonomy captures many issues associated with data typically used in empirical software engineering modeling, although some of the elements of the taxonomy are not peculiar to ESE data sets. The issues are grouped into three main classes. First is the group of characteristics of data that mean the

observations are not fit for model-building (accuracy); second are data set characteristics that lead to concerns about the suitability of applying one model to another data set (relevance); and third is a set of factors that limit data accessibility and trust (provenance) [7]. Fig. 4 depicts the ESE taxonomy.

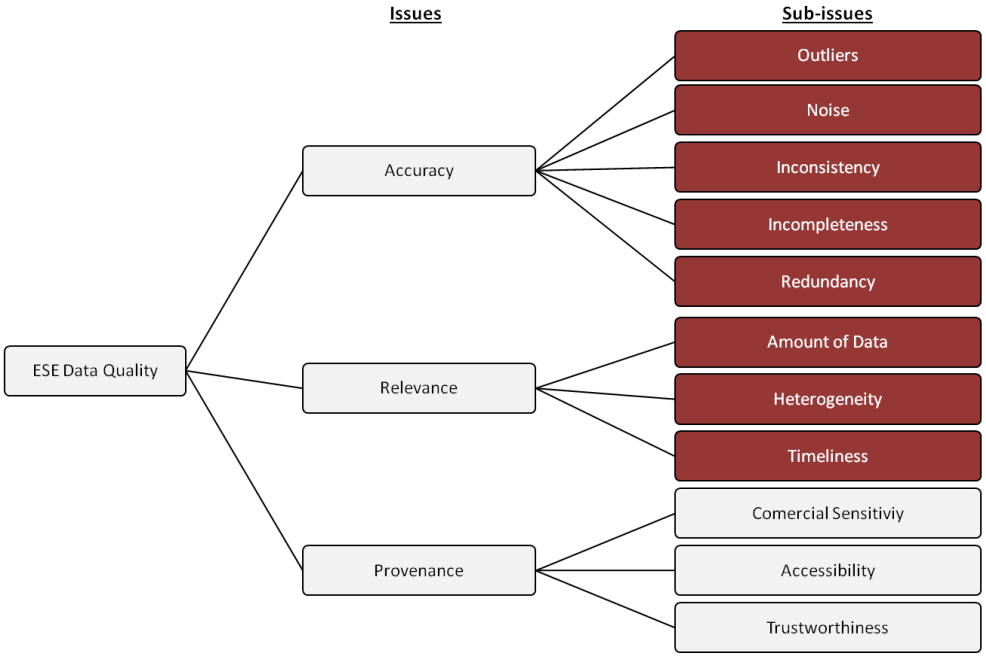

Fig. 4. Taxonomy of data quality challenges in empirical software engineering (ESE).

We took all sub-issues of accuracy and relevance (Fig. 5) as starting point for the construction of FDQ-KDT framework proposed. The provenance sub-issues not were considered because we assume the data has availability.

3. Framework for Data Quality in Knowledge Discovery Tasks (FDQ-KDT)

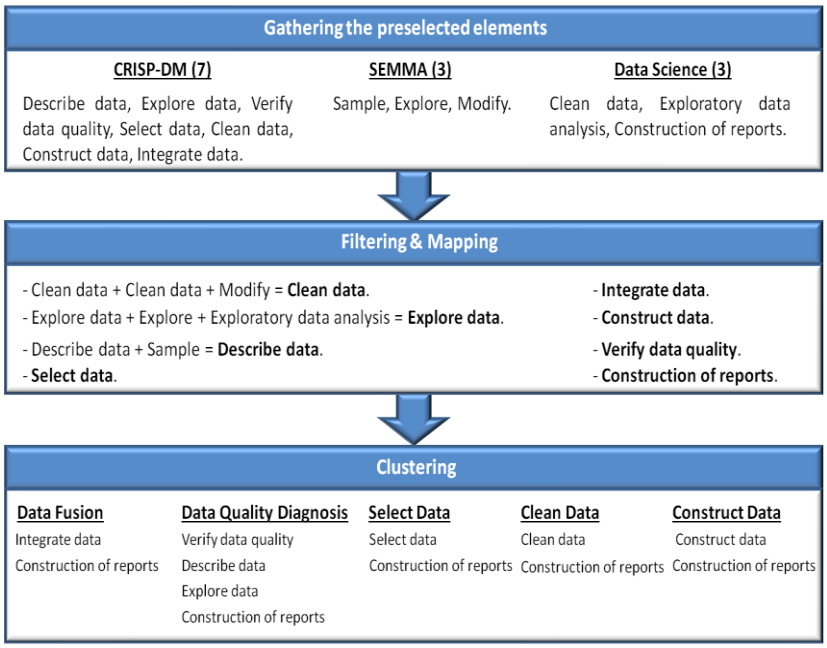

Fig. 5. Flow of development for FDQ-KDT.

The proposed framework was developed to address poor quality data in knowledge discovery task such

as data mining and machine learning projects. Fig. 5 shows the process of developing the proposed framework based on methodology developed by Almutiry [12]. The process began with *Gathering the preselected elements* of CRISP-DM, SEMMA and Data science area (red color elements in Fig. 1, Fig. 2, and Fig. 3). Afterward in *Filtering & Mapping Phase* (Fig. 5) the repeated components were removed. The *Clustering phase*, grouped the remaining components in five phases: data fusion, data quality diagnosis, select data, clean data, and construct data.

The result to apply the methodology of Almutiry [12], is the FDQ-KDT which comprising seven tasks of the phases of data understanding and data preparation of CRISP-DM (*describe data, explore data, verify data quality, select data, clean data, construct data*, and *integrate data*), three stages of SEMMA (*modify, sample, explore*), and three steps of Data Science Area (*clean data, exploratory data analysis* and *construction of reports*), organized in five main phases (Fig. 5).

In this regard, the FDQ-KDT phases have an execution order with the aim of supply a tidy dataset. Fig. 6 shows the execution process of FDQ-KDT.

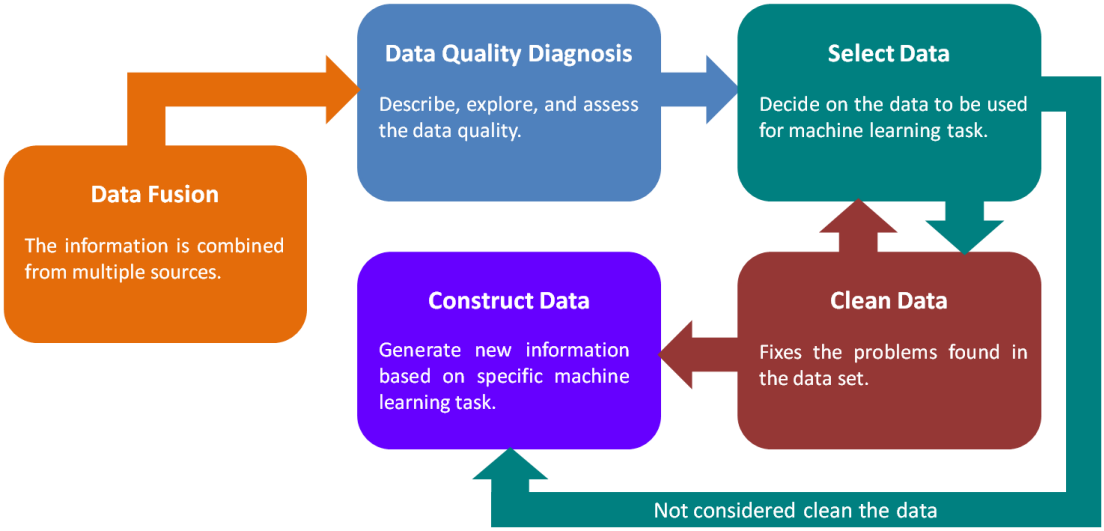

Fig. 6. Execution process of FDQ-KDT.

Fig. 6 shows the execution process of proposed framework. The process begins by combining information of multiple sources in data fusion component with aim to create a dataset (if the information comes of one source, this step is avoided). Subsequently is assessed the quality of new dataset in data quality diagnosis component. The next step selects the data more suitable for knowledge discovery task taking into account selection criteria such as data quality diagnosis, among others. If diagnosis of dataset is good, it will be sent to construct data component for completing the execution process of framework otherwise to clean data component for fixing the problem found.

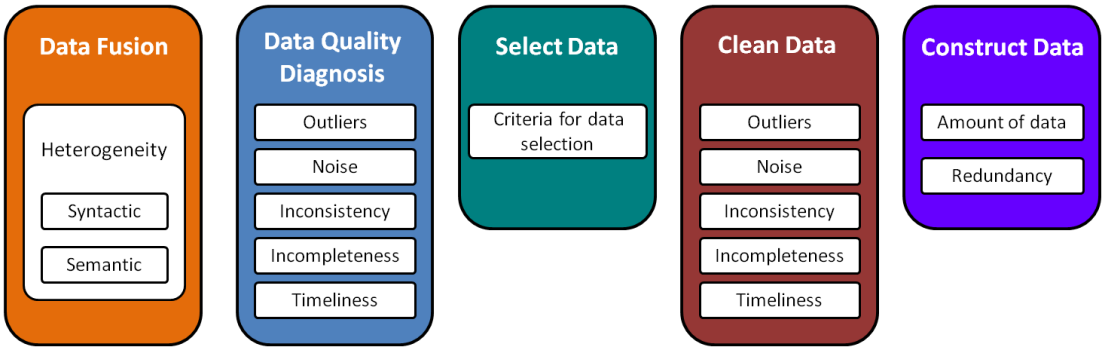

Fig. 7. Components of FDQ-KDT.

The issues of relevance and accuracy (Fig. 4 red branches) of Taxonomy of Data Quality Challenges in Empirical Software Engineering are embedded in the five components of FDQ-KDT. The basic idea is to resolve the relevance and accuracy issues found in ESE taxonomy for any domain through FDQ-KDT. Fig. 7 shows the ESE taxonomy issues organized in the five phases of FDQ-KDT.

Next is presented a detailed description of framework components for data quality in knowledge discovery tasks:

3.1. Data Fusion

Better known as *Integrate data* in CRISP-DM, involves combining information from multiple sources such as databases, plain text, physical document, etc. Data fusion encourages the issue *heterogeneity* defined as incompatibility of information. We distinguish two types of heterogeneity: s*yntactic heterogeneity* refers to differences among definitions, such as attribute types, formats, or precision, while *semantic heterogeneity* refers to differences or similarities in the meaning of data [13].

Nowadays several techniques exist for solving the *heterogeneity* issue through data fusion approaches. In [14] the authors propose three categories of data fusion techniques: data association, state estimation, and decision fusion, as shown in Table 1.

Table 1. Data Fusion Techniques

Category	Techniques
Data association	Nearest Neighbors and K-Means, Probabilistic Data Association, Joint Probabilistic Data Association, Multiple Hypothesis Test, Distributed Joint Probabilistic Data Association, Distributed Multiple Hypothesis Test, Graphical Models.
State estimation	Maximum Likelihood and Maximum Posterior, Kalman Filter, Particle Filter, Distributed Kalman Filter, Distributed Particle Filter, Covariance Consistency Methods: Covariance Intersection/Union, Covariance Union.
Decision fusion	Bayesian Methods, Dempster-Shafer Inference, Abductive Reasoning, Semantic Methods.

3.2. Data Quality Diagnosis

This component gather the tasks: *description data (sample in* SEMMA*)*, *exploration data (explore in* SEMMA *and exploratory data analysis in* Data Science*)* and *verification of data quality* of CRISP-DM. *Description data* examines the "gross" or "surface" properties of the acquired data and reports on the results, examining issues such as the format of the data, the quantity of the data, the number of instances and attributes, the identities of the attributes, and any other surface features of the data. *Exploration data* eliminates or sharpens potential hypotheses about the world that can be addressed by the data, through statistical analyses. These analyses may directly address the knowledge discovery goals; they may also contribute to or refine the data description and quality reports [11]. *Verification of data quality* assess the quality of the data, considering the next issues (Fig. 7):

- **Outliers:** these are observations which deviate so much from other observations as to arouse suspicions that it was generated by a different mechanism [15]. Outlier detection is used extensively in many applications. Current application areas of outlier detection include detection of credit card frauds, detecting fraudulent applications or potentially problematic customers in loan application processing, intrusion detection in computer networks, medical condition monitoring such as heart-rate monitoring, identifying abnormal health conditions, detecting abnormal changes in stock prices and fault diagnosis [16].

- **Noise:** defined as irrelevant or meaningless data [17] in the instances. For a given domain-specific dataset, attributes that contain a significant amount of noise can have a detrimental impact on the success of a knowledge discovery initiative, e.g., reducing the predictive ability of a classifier in a supervised learning task [18].
- **Inconsistency:** refers to a lack of harmony between different parts or elements; instances that are self-contradictory, or lacking in agreement when it is expected [7]. This problem is also known as mislabeled data or class noise. e.g., in supervised learning tasks, two instances have the same values, but have different labels or the label values do not correspond itself.
- **Incompleteness:** it is widely acknowledged as data sets affected by missing values. Typically occur because of sensor faults, a lack of response in scientific experiments, faulty measurements, data transfer problems in digital systems or respondents' unwillingness to respond to survey questions [19].
- **Timeliness:** has been defined as the degree to which data represent reality from the required point in time. When the state of the world changes faster than our ability to discover these state changes and up-date the data repositories accordingly, the confidence on the validity of data decays with time [20]. e.g., people move, get married, and even die without filling out all necessary forms to record these events in each system where their data is stored [21].

3.3. Select Data

Decide on the data to be used for analysis. Criteria include relevance to the knowledge discovery task, the assessment performed in *data quality diagnosis component (DQD)* and technical constraints such as limits on amount of data or redundancies (*construct data component* in Section 3.5). If diagnosis of dataset is good, it will be sent to construct data component otherwise to clean data component for fix the problem found (Fig. 6).

3.4. Clean Data

Known in CRISP-DM *and Data Science as Clean Data and* SEMMA *as Modify*, improved the data quality with regard to issues found in *DQD*. This may involve selection of clean subsets of the data, the insertion of suitable defaults [11], or another approaches to fix the issue such as outliers, noise, inconsistency, incompleteness, timeliness (Fig. 7).

In Table 2 we presented the approaches to fix the issues found in *DQD phase* from the review works [18], [22]–[24] and the research [25].

Table 2. Approaches to Fix the Issues Found in *DQD Phase*

Issues	Approaches	Research
Outliers	Statistical, Nearest Neighbor, Clustering, Classification, Spectral Decomposition.	[22]
Noise	Statistical, Nearest Neighbor, Clustering, Classification, Information Theoretic Spectral.	[18]
Inconsistency	Classification, Ensemble methods, Clustering, Probabilistic methods.	[23]
Incompleteness	Imputation methods: Mean-Mode, Hotdeck, K-nearest neighbor, Multiple, Multivariate by chained equations, Clustering.	[24]
Timeliness	Supervised learning for information-decay-based predictive models	[25]

3.5. Construct Data

This component builds a dataset for a specific task of knowledge discovery (classification, clustering,

etc.), also generates new information from that production of derived attributes or transformed values for existing attributes, solving the issues (Fig. 7):

- **Amount of data:** the amount of data available for model building contributes to relevance in terms of goal attainment [7]; small and imbalanced data sets build inaccurate models.
- **Redundancy:** as the name implies is the redundant information such as duplicate instances and derived attributes of others that contain the same information [26], [27].

In Table 3 are shown the approaches to solve the problems related to amount of data and redundancy.

Table 3. Approaches to Fix the Issues: Amount of Data and Redundancy

Issues	Approaches	Research
Amount of data	Synthetic dataset, Resampling of imbalanced datasets, Incremental learning.	[28]–[30]
Redundancy	Feature and instance selection methods: wrapper, filter.	[26], [27]

It is worth mentioning that construction of reports (step of SEMMA methodology) is taking into account in all phases of FDQ-KDT.

4. Conclusion and Future Research

In knowledge discovery tasks such as classification, prediction, cluster, etc, is very important to use tidy dataset to get relevant outcomes. In the early decades of computing, a common saying was "garbage in, garbage out." That is, mistakes in recollection of information were aberrations, and if knowledge discovery tasks have bad data (garbage in), then they should expect incorrect answers (garbage out) [31]. For this reason we proposed a conceptual framework for data quality in knowledge discovery task based on CRISP-DM, SEMMA and Data Science Area, which tackle the issues in data quality clearly through ESE taxonomy.

Several approaches exist to tackle the issues of data quality in outliers [22], noise [18], inconsistency [23], incompleteness [24], redundancy [26], [27], amount of data [28]–[30], heterogeneity [14], and timeliness [25]. Nevertheless the results to date not consider resolve the issues in ensemble. Thus the next step will be developing, examining and evaluating the proposed framework through artificial intelligence algorithms, statistical and mathematical models.

Acknowledgements

The authors are grateful to the Telematics Engineering Group (GIT) of the University of Cauca, Control Learning and Systems Optimization Group (CAOS) of the Carlos III University of Madrid for technical support and Colciencias for PhD scholarship granted to MsC. David Camilo Corrales.

References

[1] Wang, L., Zhan, J., Luo, C., Zhu, Y., Yang, Q., He, Y., Gao, W., Jia, Z., Shi, Y., Zhang, S., Zheng, C., Lu, G., Zhan, K., Li, X., & Qiu. B., (2014). BigDataBench: A big data benchmark suite from internet services. *Proceedings of 2014 IEEE 20th International Symposium on High Performance Computer Architecture (HPCA)* (pp. 488–499).

[2] Rajaraman, A., & Ullman, J. D. (2011). *Mining of Massive Datasets*. New York, N.Y.; Cambridge: Cambridge University Press.

[3] Pacheco, F., Rangel, C., Aguilar, J., Cerrada, M., & Altamiranda, J. (2014). Methodological framework for data processing based on the data science paradigm. *Proceedings of 2014 XL Latin American Computing Conference (CLEI)* (pp. 1–12).

[4] Olson, D. L., & Delen, D. (2008). *Advanced Data Mining Techniques* (1st ed.). Springer Publishing Company, Incorporated.

[5] Azevedo, A. I. R. L. (2008). KDD, SEMMA and CRISP-DM: A parallel overview. *Proceedings of IADIS European Conf. Data Mining* (pp. 182–185).

[6] O'Neil, C., & Schutt, R. (2013). *Doing Data Science: Straight Talk from the Frontline*, Edición: 1. O'Reilly Media.

[7] Bosu, M. F., & MacDonell, S. G. (2013). A Taxonomy of data quality challenges in empirical software engineering. *Proceedings of 2013 22nd Australian Software Engineering Conference (ASWEC)* (pp. 97–106).

[8] Kerr, K., & Norris, T. (2014). The development of a healthcare data quality framework and strategy. *Proceedings of the Ninth International Conference on Information Quality (ICIQ-04)* (pp. 218–233).

[9] Wang, R. Y., & Strong, D. M. (1996). Beyond accuracy: What data quality means to data consumers. *J Manage Inf Syst*, *12(4),* 5–33.

[10] Wittig, D. (2000). Conceptualizing information quality: A review of information quality frameworks from the last ten years. *Proceedings of the 2000 Conference on Information Quality* (pp. 83–96).

[11] Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirt, R., (1999). *CRISP-DM 1.0 Step-by-Step Data Mining Guide*.

[12] Almutiry, O., Wills, G., Alwabel, A., Crowder, R., & WaIters, R. (2013). Toward a framework for data quality in cloud-based health information system. *Proceedings of 2013 International Conference on Information Society (i-Society)* (pp. 153–157).

[13] Hakimpour, F., & Geppert, A. (2001). Resolving semantic heterogeneity in schema integration. *Proceedings of the International Conference on Formal Ontology in Information Systems* (pp. 297–308). New York, NY, USA.

[14] Castanedo, F. (2013). A review of data fusion techniques. *Sci. World J.*, e704504.

[15] Hawkins, D. M. (1980). *Introduction, in Identification of Outliers*, 1–12, Springer Netherlands.

[16] Daneshpazhouh, A., & Sami, A. (2014). Entropy-based outlier detection using semi-supervised approach with few positive examples. *Pattern Recognit. Lett.*, *49,* 77–84.

[17] Xiong, H., Pandey, G., Steinbach, M., & Kumar, V., (2006). Enhancing data analysis with noise removal. *IEEE Trans. Knowl. Data Eng.*, *18(3)*, 304–319.

[18] Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly detection: A survey. *ACM Comput. Surv. CSUR*, *41(3)*, 15.

[19] Aydilek, I. B., & Arslan, A. (2013). A hybrid method for imputation of missing values using optimized fuzzy c-means with support vector regression and a genetic algorithm. *Inf. Sci.*, *233,* 25–35.

[20] Finger, M., & Da, S. F. S. (1998). Temporal data obsolescence: Modelling problems. *Proceedings of Fifth International Workshop on Temporal Representation and Reasoning* (pp. 45–50).

[21] Maydanchik, A. (2007). *Data Quality Assessment*. Technics Publications.

[22] Shukla, D. S., Pandey, A. C., & Kulhari, A. (2014). Outlier detection: A survey on techniques of WSNs involving event and error based outliers. *Proceedings of 2014 Innovative Applications of Computational Intelligence on Power, Energy and Controls with Their Impact on Humanity (CIPECH)* (pp. 113–116).

[23] Frenay, B., & Verleysen, M. (2014). Classification in the presence of label noise: A survey. *IEEE Trans. Neural Netw. Learn. Syst.*, *25(5)*, 845–869.

[24] Thirukumaran, S., & Sumathi, A. (2012). Missing value imputation techniques depth survey and an imputation algorithm to improve the efficiency of imputation. *Proceedings of 2012 Fourth International Conference on Advanced Computing (ICoAC)* (pp. 1–5).

[25] Placide, M., & Lasheng, Y. (2010). Information decay in building predictive models using temporal data.

Proceedings of 2010 International Symposium on Information Science and Engineering (ISISE) (pp. 458–462).

[26] Khalid, S., Khalil, T., & Nasreen, S. (2014). A survey of feature selection and feature extraction techniques in machine learning. *Science and Information Conference (SAI)* (pp. 372–378).

[27] Olvera-López, J. A., Carrasco-Ochoa, J. A., Martínez-Trinidad, J. F., & Kittler, J. (2010). A review of instance selection methods. *Artif. Intell. Rev.*, *34(2)*, 133–143.

[28] Tomás, J. T., Spolaôr, N., Cherman, E. A., & Monard, M. C. (2014). A framework to generate synthetic multi-label datasets. *Electron. Notes Theor. Comput. Sci.*, *302*, 155–176.

[29] Cateni, S., Colla, V., & Vannucci, M. (2014). A method for resampling imbalanced datasets in binary classification tasks for real-world problems. *Neurocomputing*, *135*, 32–41.

[30] Kidera, T., Ozawa, S., & Abe, S. (2006). An Incremental learning algorithm of ensemble classifier systems. *Proceedings of International Joint Conference on Neural Networks* (pp. 3421–3427).

[31] Phalgune, A., Kissinger, C., Burnett, M., Cook, C., Beckwith, L., & Ruthruff, J. R. (2005). Garbage in, garbage out? An empirical look at oracle mistakes by end-user programmers. *Proceedings of 2005 IEEE Symposium on Visual Languages and Human-Centric Computing* (pp. 45–52).

David Camilo Corrales received the degree in informatics engineering and master degree in telematics engineering at University of Cauca, Colombia, in 2011 and 2014 respectively. Actually he is a PhD student in telematics engineering at the University of Cauca and Science and Informatics Technologies at Carlos III of Madrid University. His research interests focus on data mining, machine learning and data analysis.

Agapito Ledezma is an associate professor in the Department of Computer Science at Carlos III of Madrid University. He received a B.S. degree from Universidad Latinoamericana de Ciencia y Tecnología in 1997. He received his Ph.D. degree in computer science from Carlos III University in 2004. His research interests center on machine learning, activity recognition, intelligent agents and advanced driving assistant systems. He has published over 80 journal and conference papers mainly in the field of artificial intelligence and machine learning.

Juan Carlos Corrales received the Dipl-Ing and master's degrees in telematics engineering from the University of Cauca, Colombia, in 1999 and 2004 respectively, and the PhD degree in sciences, specialty computer science, from the University of Versailles Saint-Quentin-en-Yve-lines, France, in 2008. Presently, he is a full professor and leads the Telematics Engineering Group at the University of Cauca. His research interests focus on service composition and data analysis.