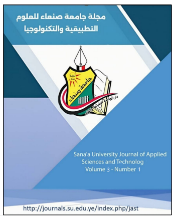
Sun Publisher

مجلة جامعة صنعاء للعلوم التطبيقية والتكنولوجيا
Sana'a University Journal of Applied Sciences and Technology
https://journals.su.edu.ye/index.php/jast/



Big Data Formation, Reduction, and Its Impact on Sampling: A survey

Mohmmed Mohammed Zayed [1*] and Fadl Mutaher Ba-Alwi [2]

[1] **Department of Computer Science, Faculty of Information Technology and computer , University of Sana'a, Sana'a, Yemen,**
[2] **Department of Information System, Faculty of Information Technology and computer , University of Sana'a, Sana'a, Yemen**

*Corresponding author: mzayed@su.edu.ye

Abstract

The emergence of data in recent years, characterized by the "6Vs" (Volume, Velocity, Variety, Veracity, Value, and Variability), has started the era of big data. While this data holds great potential for uncovering valuable insights and knowledge, its size presents significant challenges for analysis. This paper explores two critical big data reduction techniques: feature selection and sampling. Feature selection focuses on identifying and eliminating irrelevant or redundant features, reducing data dimensionality. Sampling, on the other hand, selects a representative subset of data points for analysis. We compare and contrast these techniques, highlighting their strengths and weaknesses. The paper explores when each approach is most suitable and suggest the potential benefits of combining them for even more efficient big data analysis.

Article Info

Keywords:
Big data, Sampling, Big data reduction, Impact of sampling.

Article History:
Received: 11-January-2025,
Revised: 20-January-2025,
Accepted: 18-February-2024,
Available online: 28 February 2025.

1. INTRODUCTION

The ever-growing volume, velocity, and variety of data we generate define the era of big data. This data originates from various sources, including social media, sensor networks, financial transactions, and scientific research. While this data holds immense potential for uncovering valuable insights, its sheer size presents significant challenges for storage, processing, and analysis these challenging brings new burdens when dealing with it. This paper explores the formation of big data, techniques for big data reduction, and the subsequent impact on sampling methodologies.

This paper aims to:

1. Explore the formation of big data, characterized by its "6Vs" (Volume, Velocity, Variety, Veracity, Value, and Variability), and the challenges they present for data analysis.
2. Investigate key big data reduction techniques, including feature selection and sampling, to understand their role in managing large datasets.
3. Analyze the impact of data reduction on sampling strategies, focusing on scalability, representativeness, and computational efficiency.
4. Compare and contrast feature selection and sampling methods to highlight their strengths, limitations, and areas of application.
5. Propose a hybrid approach combining feature selection and sampling for optimal big data reduction.
6. Suggest future directions for advancing big data reduction techniques to address real-world analytical challenges.

2. BACKGROUND

2.1. Formation of Big Data

Big data is characterized by the "6Vs":

- **Volume:** The sheer amount of data generated is massive and constantly growing.
- **Velocity:** Data is generated and collected at an unprecedented rate, requiring real-time or near real-time processing.
- **Variety:** Big data comes in various formats, including

structured (databases), semi-structured (logs), and unstructured (text, images, videos).
- **Veracity:** Data quality can be an issue, with inconsistencies, errors, and missing values requiring careful cleaning and further validation.
- **Value:** Extracting meaningful insights from the vast amount of data is crucial to unlock its potential value.
- **Variability:** Data can change rapidly, requiring adaptable and scalable systems for continuous analysis [1].

These characteristics create a complex data landscape that traditional data management techniques struggle or even fail to handle.

2.2. Sources of Big Data

- **Social media**: Platforms like Facebook, Twitter, and Instagram generate vast amounts of user data.
- **IoT Devices**: Sensors and devices connected to the internet provide continuous streams of data.
- **Transactional Data**: Online shopping, banking, and other services record transactions continuously.
- **Multimedia Data**: Images, videos, and audio files uploaded to platforms like YouTube and Spotify.
- **Geospatial Data**: Data from GPS devices and location-based services.

2.3. Big Data Reduction

Big data reduction is the process of minimizing the volume of data without losing its essential characteristics and insights. Techniques for data reduction include:

1. **Data Compression:** Reduces the size of the data files using algorithms that encode information more efficiently. Techniques include lossless compression (e.g., ZIP) and lossy compression (e.g., JPEG for images) [2].
2. **Data Aggregation:** Summarizes detailed data into a more manageable form. For instance, instead of storing every transaction, daily or monthly summaries can be used [1].
3. **Dimensionality Reduction:** Involves reducing the number of variables under consideration. Techniques such as Principal Component Analysis (PCA) and t-Distributed Stochastic Neighbor Embedding (t-SNE) useful in reducing the complexity of data [3].
4. **Sampling:** Selecting a representative subset of the data to make inferences about the entire dataset. Sampling techniques include random sampling, stratified sampling, and systematic sampling [4, 5, 6].
5. **Data Filtering:** Removes irrelevant or redundant data. Techniques such as outlier detection and noise reduction are employed to clean the data [7].

3. MPIRICAL ANALYSIS AND METHODOLOGY

3.1. Impact on Sampling

Sampling is a critical process in big data analytics, providing a means to make data analysis feasible and cost-effective thus can be handled. The formation and reduction of big data significantly impact sampling strategies. Big data reduction techniques significantly impact the way sampling is conducted:

- **Targeted Sampling:** Reduction techniques like data aggregation can reveal patterns and trends, allowing for more targeted sampling approaches focused on specific sub-populations in the data [5].
- **Scalable Sampling Techniques:** Big data frameworks like Apache Spark enable efficient sampling of massive datasets, making it possible to draw statistically significant samples even from very large datasets [1].
- **Data Quality Considerations:** Reduction techniques like filtering and cleaning can inadvertently introduce bias into the data. Careful consideration of data quality is crucial to ensure the representativeness of the sampled data [7].
- **New Sampling Techniques:** The unique characteristics of big data, such as real-time data streams, necessitate the development of new sampling techniques like stream sampling, which can extract representative subsets from continuously flowing data [3, 8].

3.2. Challenges in Sampling Big Data

Sampling in big data presents new challenges, as traditional sampling methods often don't perform well due to the volume, velocity, and variety

inherent in large datasets. Unlike conventional data, big data is unstructured, originating from many sources such as social media, sensors, and transactional logs, which harden the selection of representative samples. Ensuring sampling validity while managing computational constraints is very important, as standard methods can be computationally expensive or infeasible when handling very big data volumes [9]. Additionally, high-dimensional data introduces new issues with data sparsity, potentially skewing results and limiting generalizability [10]. Sampling biases are also more likely in big data environments due to unobserved heterogeneity, dynamic changes in data streams, and uneven distribution across features, all of which can threaten the representativeness of the sample and lead to inaccurate inferences [11]. Addressing these challenges requires innovative sampling methods deigned to big data's scale and complexity to maintain analytical reliability and ensure data-driven insights remain valid with low computational cost, the following highlighters main challenges factors in big data sampling:

1. **Representativeness**: Ensuring that the sample accurately reflects the population is challenging given the variety and volume of data.
2. **Scalability**: Traditional sampling methods may not scale efficiently with the increasing size of datasets.
3. **Bias and Variance**: Bias and variance are two fundamental sources of error in machine learning models: bias refers to the error from simplifying assumptions in the model, leading to underfitting, while variance denotes the error due to the model's sensitivity to small fluctuations in the training set, often resulting in overfitting. High-dimensional data can exacerbate these issues due to the "curse of dimensionality," where sparse data distributions make it difficult to form reliable patterns, increasing variance as the model fits to noise and outliers [12]. Additionally, high-dimensional spaces can introduce sampling bias, where certain regions of the data space are underrepresented, leading models to make overly simplistic assumptions, increasing bias [13]. Techniques such as dimensionality reduction (e.g., Principal Component Analysis) and regularization (e.g., Ridge and Lasso) have been used to mitigate these issues by reducing model complexity, thereby balancing the bias-variance tradeoff in high-dimensional settings [14].

3.3. Strategies for Effective Sampling

The sample is a specific group from which data is collected, ideally representing a larger population of interest. Sampling methods are generally classified into probability sampling and non-probability sampling. Probability sampling, where each member of the population has an equal chance of selection, enhances the generalizability of findings and includes techniques like simple random sampling, stratified sampling, cluster sampling, and systematic sampling [15]. Non-probability sampling, in contrast, involves non-random selection methods, which can introduce biases but are often more feasible for exploratory or resource-constrained studies. Common non-probability methods include convenience sampling, quota sampling, purposive sampling, and snowball sampling [16]. Selecting an appropriate sampling technique is essential, as each method has unique strengths and limitations, impacting data validity and research conclusions.

3.3.1. Probability Sampling Methods

Probability sampling methods are foundational techniques in research for ensuring that each individual in a population has a known and non-zero chance of selection, allowing for unbiased and generalizable results. These methods, which include random sampling, stratified sampling, cluster sampling, and systematic sampling, are widely valued for their ability to produce representative samples while minimizing selection bias and enhancing the validity of statistical inferences [17]. For example, stratified sampling divides the population into distinct subgroups, or strata, and samples within each, increasing precision by capturing variations across subpopulations [18]. Cluster sampling is particularly useful for large populations spread across vast areas, as it selects groups, or clusters, rather than individuals, thereby reducing sampling costs while maintaining representativeness [19]. Each probability sampling method has specific advantages and limitations, and choosing the appropriate method depends on the research goals, population characteristics, and available resources as follows:

1. **Stratified Sampling**: Divides the population into strata and samples each stratum proportionally. This ensures that different segments of the data are adequately represented.
2. **Systematic Sampling**: Selects every k-th item from a sorted list. This method is useful when data is ordered or has a periodic structure.
3. **Cluster Sampling**: Divides the data into clusters and randomly selects clusters to sample from. This method reduces the cost and complexity of data collection.
4. **Adaptive Sampling**: Adjusts the sampling technique based on real-time data analysis, which can be useful in dynamic environments where data characteristics change rapidly.
5. **Importance Sampling**: Focuses on selecting data points that have higher significance or weight. This method is often used in machine learning to improve model training efficiency.

3.3.2. Non-Probability Sampling Methods

Non-probability sampling methods are approaches in which individuals are selected from the population without the principle of randomization, meaning each member does not have a known or equal chance of being included. These methods are often employed when probability sampling is impractical due to constraints such as limited resources, access challenges, or the nature of exploratory research where generalizability is not the primary focus.

While non-probability sampling methods are less robust in terms of producing statistically generalizable results, they can be valuable in generating initial insights, understanding specific groups, or collecting data in contexts where randomness is not feasible [20]. Types of Non-Probability Sampling are as follows:

1. **Convenience Sampling**
   Convenience sampling is a widely used method in which the researcher selects participants based on ease of access and availability. Often applied in early-

stage research or exploratory studies, convenience sampling is simple and cost-effective. However, its lack of randomness introduces selection bias, and the sample may not accurately represent the larger population [21]. For instance, collecting survey responses from people passing through a specific location limits the diversity and may skew results toward certain demographics.

2. **Quota Sampling**
Quota sampling involves segmenting the population into exclusive subgroups and then selecting a predefined number of participants from each group. This approach ensures that the sample includes specific proportions of the population's characteristics, such as age, gender, or socioeconomic status, to capture relevant diversity [22]. However, since participants within each quota are often chosen based on availability rather than random selection, quota sampling may still suffer from selection bias, potentially limiting the sample's representativeness [23].
3. **Purposive (or Judgmental) Sampling**
Purposive sampling, also known as judgmental sampling, involves the researcher selecting participants based on specific criteria or characteristics relevant to the research question. This method is commonly used in qualitative research where the goal is to gather in-depth insights from a particular group, such as experts in a field or individuals with a specific experience [24]. While purposive sampling enables the researcher to target highly relevant individuals, it lacks generalizability and is subject to researcher bias, as participant selection is based on subjective judgment.
4. **Snowball Sampling**
Snowball sampling is a method often used when studying hard-to-reach or hidden populations, such as marginalized communities or individuals with specific behaviors. In this method, initial participants, known as "seeds," recruit additional participants from among their acquaintances, creating a chain referral process [25]. Snowball sampling is valuable for accessing niche groups, but it is prone to bias, as the sample may reflect the network characteristics of the initial participants rather than the entire population.
5. **Judgmental Sampling**
This technique involves the researcher's deliberate selection of participants based on their knowledge or expertise, particularly in cases where specialized insights are necessary. Common in fields requiring expert opinions, such as policymaking or medical research, judgmental sampling allows for targeted data collection from knowledgeable sources. However, its reliance on subjective judgment means that findings may not extend beyond the sampled individuals, limiting generalizability [26].

3.4. Feature Selection

Feature selection involves identifying and retaining the most relevant features (variables) from the dataset. This process reduces the dimensionality of the data, leading to faster processing times, less storage space, and improved model performance [27].

- **Importance of Feature Selection**

1. **Enhanced Model Performance**: By eliminating noise and redundant information, models perform better.
2. **Reduced Overfitting**: Fewer features help reduce the risk of overfitting, where models perform well on training data but poorly on new data.
3. **Lower Computational Cost**: Reducing the number of features decreases the computational resources needed for training and predicting.
4. **Improved Interpretability**: Models with fewer features are easier to understand and interpret.

- **Techniques for Feature Selection:**

Feature selection techniques can be broadly categorized into the following:

1. **Filter Methods**: Evaluate the relevance of features based on intrinsic properties, independently of any machine learning algorithm. Common techniques include correlation coefficients, chi-square tests, mutual information, and variance thresholds.
2. **Wrapper Methods**: Assess the usefulness of feature subsets by training models on them and evaluating performance. Techniques include forward selection, backward elimination, and recursive feature elimination (RFE).
3. **Embedded Methods**: Perform feature selection during model training, specific to certain algorithms. Examples include LASSO (Least Absolute Shrinkage and Selection Operator) and tree-based methods like Random Forests.
4. **Hybrid Methods**: Combine the strengths of filter and wrapper methods. They often use a filter method to initially reduce the feature set and then apply a wrapper method on the reduced set for further refinement.
5. **Heuristic-Based Methods**: These methods use heuristic search strategies, such as genetic algorithms, particle swarm optimization, or simulated annealing, to find optimal subsets of features.
6. **Laplacian Score-Based Feature Selection**: A method from manifold learning, Laplacian scores select features by evaluating their local discriminative power. This method is useful when trying to preserve the intrinsic structure of data, especially in nonlinear relationships.
7. **Information-Theoretic Approaches**: These methods select features based on information-theoretic

measures such as entropy, mutual information, or information gain, aiming to maximize the information retained in the selected features.

8. **Sparse Learning-Based Methods:** Sparse learning methods impose sparsity constraints, such as L1 regularization, to force the model to select only a few features by shrinking the coefficients of less important features to zero.
9. **Embedded Deep Learning Methods**: Some deep learning models, especially neural networks, can implicitly perform feature selection as part of the training process. Techniques such as feature importance ranking in neural networks or attention mechanisms in transformers serve to highlight important features.
10. **Principal Feature Analysis (PFA)** : PFA is an extension of PCA (Principal Component Analysis) that identifies and selects the most representative features based on principal components, retaining interpretability while reducing dimensionality.

4. COMPARING FEATURE SELECTION AND SAMPLING IN BIG DATA REDUCTION

As data volume and dimensionality continue to grow, feature selection and sampling have become crucial in making big data manageable for analysis and modeling. This section provides a detailed comparative analysis of feature selection and sampling, explores their situational strengths and limitations, and introduces empirical scenarios to inform best practices for practitioners.

4.1. Feature Selection in Big Data

Feature selection reduces the number of variables in a dataset, focusing on the most relevant features for analysis. This technique addresses high-dimensionality challenges by removing irrelevant or redundant information, which can improve model accuracy and processing efficiency [28].

- **Advantages of Feature Selection:**

1. **Enhanced Model Performance**: By removing irrelevant features, models can focus on the most important variables, leading to better performance.
2. **Reduced Overfitting**: Fewer features mean less noise, which helps in reducing overfitting.
3. **Lower Computational Cost**: Reducing the number of features decreases the resources required for data processing and model training.
4. **Improved Interpretability**: Models with fewer, more meaningful features are easier to understand and interpret.

- **Disadvantages**

1. **Complex Interdependencies**: It can be challenging to capture complex relationships between features.
2. **Scalability Issues**: Feature selection methods may become computationally intensive with extremely high-dimensional data.

- **Empirical Scenario**

In high-dimensional datasets, such as genetic data (where each feature could represent a genetic marker), feature selection techniques like Principal Component Analysis (PCA) and Recursive Feature Elimination (RFE) can reduce thousands of variables to a manageable number. For instance, PCA effectively condenses the dataset by projecting it into a lower-dimensional space, while preserving variance, which improves model performance without overfitting.

4.2. Sampling

Sampling reduces data volume by selecting representative known as subset from the full dataset. Unlike feature selection, sampling keeps all variables intact, allowing models to be built with less data, which is particularly useful for extremely large datasets or real-time data streams.

- **Advantages of Sampling:**

1. **Scalability**: Sampling can handle very large datasets, making it a practical solution for big data.
2. **Reduced Computational Cost**: By analyzing a smaller subset, resources are conserved.
3. **Feasibility**: Makes it possible to perform data analysis when the full dataset is too large to handle.

- **Disadvantages:**

1. **Representativeness**: Ensuring that the sample accurately reflects the entire dataset can be challenging.
2. **Bias and Variance**: Poor sampling techniques can introduce bias and affect the variance, leading to inaccurate conclusions.
3. **Data Loss**: Important information might be missed if the sample is not representative.

- **Empirical Scenario**

In real-time streaming data, such as data from IoT sensors, systematic sampling techniques (like stream sampling) help capture representative data subsets on the fly. For instance, Apache Spark's sampling functions can draw statistically significant samples even from rapidly changing data streams, balancing between computational cost and real-time processing needs [29].

4.3. Combined Approach for Optimal Big Data Reduction

In many real-world applications, combining feature selection and sampling techniques can lead to improved performance and efficiency. A hybrid approach is especially beneficial in situations with high-dimensional and high-volume data, where both volume and dimensionality need reduction for effective analysis [30].

- **Empirical Scenario for Combined Approach**

Consider a large social media dataset used for sentiment analysis, where the data is both high-dimensional (text data) and vast in volume. First, feature selection is applied (e.g., term frequency-inverse document frequency, TF-IDF), it reduces the dimensionality by focusing on the most relevant keywords or phrases. Then, sampling can further reduce the data size, making it feasible to analyze the data in real time.

By using this combined approach, the dataset is reduced in both dimensionality and volume, facilitating faster processing while preserving essential patterns for analysis.

4.4. Quantitative Evaluation and Best Practices

To further validate these techniques, empirical analysis should be conducted on varied dataset types to compare performance under different data conditions. Here is a recommended approach:

1. **Dataset Types**: Text data (NLP tasks), image data (for image recognition), and streaming data (IoT or sensor data) should be evaluated for diverse perspectives.
2. **Performance Metrics**:
   - **Model Accuracy**: Compare accuracy with and without data reduction.
   - **Processing Time**: Measure preprocessing and training times to assess computational efficiency.
   - **Memory Usage**: Quantify memory requirements for different methods.
3. **Evaluation Results Interpretation**: Results could reveal insights such as:
   - **High-dimensional data** benefits most from feature selection.
   - **High-volume streaming data** favors sampling for faster, cost-effective analysis.
   - Combining both techniques may yield optimal results when datasets are both high-dimensional and high-volume.

4.5. Practical Recommendations

1. **When to Use Feature Selection**: Use feature selection if the primary goal is to improve model interpretability, reduce overfitting, or if dimensionality is exceptionally high.
2. **When to Use Sampling**: Sampling is preferable when computational resources are limited, or when rapid processing is required, as in real-time or streaming data applications.
3. **Combining Approaches**: Start with feature selection to eliminate irrelevant features, then apply sampling for large datasets where both dimensionality and size need reduction.

5. CONCLUSION

Feature selection reduces dimensionality by retaining only the most relevant features, leading to improved model performance, reduced overfitting, lower computational costs, and enhanced interpretability. On the other hand, sampling reduces data volume by selecting representative subsets, enabling scalable and feasible analysis, especially for massive datasets or real-time streams. The comparative analysis in this study demonstrated that both techniques are complementary and can be combined to optimize performance for high-dimensional and high-volume datasets.

The empirical scenarios presented emphasize the importance of selecting the right technique based on dataset characteristics and analysis goals. For instance, feature selection is advantageous when dealing with high-dimensional datasets, while sampling is more suitable for massive data volumes requiring fast processing. In some cases, a hybrid approach—combining feature selection and sampling—is most effective. Future work shall focus explicitly of how to combine feature selection and sampling for a better reduction of the big data.

REFERENCES

[1] Mohammad Sultan Mahmud et al. “A survey of data partitioning and sampling methods to support big data analysis”. In: *Big Data Min. Anal.* 3.2 (2020), pp. 85–101.

[2] Kamlesh Kumar Pandey and Diwakar Shukla. “Stratified sampling-based data reduction and categorization model for big data mining”. In: *Communication and Intelligent Systems: Proceedings of ICCIS 2019*. Springer Singapore, 2020, pp. 107–122.

[3] Kheyreddine Djouzi, Kadda Beghdad-Bey, and Abdenour Amamra. “A new adaptive sampling algorithm for big data classification”. In: *J. Comput. Sci.* 61 (2022), p. 101653.

[4] R. Iliyasu and I. Etikan. “Comparison of quota sampling and stratified random sampling”. In: *Biom. Biostat. Int. J. Rev* 10.1 (2021), pp. 24–27.

[5] Gaganpreet Sharma. “Pros and cons of different sampling techniques”. In: *Int. J. Appl. Res.* 3.7 (2017), pp. 749–752.

[6] Andrea E. Berndt. “Sampling methods”. In: *J. Hum. Lact.* 36.2 (2020), pp. 224–226.

[7] Tawfiq Hasanin et al. “Severely imbalanced big data challenges: investigating data sampling approaches”. In: *J. Big Data* 6.1 (2019), pp. 1–25.

[8] Kamlesh Kumar Pandey and Diwakar Shukla. “Euclidean distance stratified random sampling based clustering model for big data mining”. In: *Comput. Math. Methods* 3.6 (2021), e1206.

[9] J. Fan, F. Han, and H. Liu. “Challenges of big data analysis”. In: *National Sci. Rev.* 1.2 (2014), pp. 293–314.

[10] H. Liang, J. Wang, and Y. Yuan. “Sampling big data: Challenges and future research directions”. In: *Inf. & Manag.* 54.8 (2017), pp. 1015–1030.

[11] H. R. Varian. “Big data: New tricks for econometrics”. In: *J. Econ. Perspect.* 28.2 (2014), pp. 3–28.

[12] C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

[13] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, 2009.

[14] J. Fan and J. Lv. “A selective overview of variable selection in high-dimensional feature space”. In: *Stat. Sinica* 20.1 (2010), pp. 101–148.

[26] H. Taherdoost. “Sampling methods in research methodology; How to choose a sampling technique for research”. In: *Int. J. Acad. Res. Manag.* 5.2 (2016), pp. 18–27.

[27] Heba Al Marwai and Ghaleb H AL-Gaphari. “A hybrid Feature Selection Method Based on Binary PSO Algorithm for Microarray Data Classification”. In: *Sana’a Univ. J. Appl. Sci. Technol.* 2.4 (2024), pp. 375–380.

[28] Samuel J. Stratton. “Population research: convenience sampling strategies”. In: *Prehospital Disaster Med.* 36.4 (2021), pp. 373–374.

[29] Ahmed Sultan Alhegami and Hussein Alkhader Alsaeedi. “A framework for incremental parallel mining of interesting association patterns for big data”. In: *Int. J. Comput.* 19.1 (2020), pp. 106–117.

[15] W. G. Cochran. *Sampling Techniques*. 3rd. Wiley, 2007.

[16] I. Etikan, S. A. Musa, and R. S. Alkassim. “Comparison of convenience sampling and purposive sampling”. In: *Am. J. Theor. Appl. Stat.* 5.1 (2016), pp. 1–4.

[17] W. G. Cochran. *Sampling Techniques*. 3rd. Wiley, 2007.

[18] S. L. Lohr. *Sampling: Design and Analysis*. 2nd. Brooks/Cole, 2010.

[19] S. K. Thompson. *Sampling*. 3rd. Wiley, 2012.

[20] I. Etikan, S. A. Musa, and R. S. Alkassim. “Comparison of convenience sampling and purposive sampling”. In: *Am. J. Theor. Appl. Stat.* 5.1 (2016), pp. 1–4.

[21] M. H. Bornstein, J. Jager, and D. L. Putnick. “Sampling in developmental science: Situations, shortcomings, solutions, and standards”. In: *Dev. Rev.* 33.4 (2013), pp. 357–370.

[22] M. P. Battaglia. “Nonprobability sampling”. In: *Encyclopedia of Survey Research Methods*. Ed. by P. J. Lavrakas. Sage, 2008, pp. 524–527.

[23] P. Sedgwick. “Quota sampling”. In: *BMJ* 347 (2013), f6343.

[24] L. A. Palinkas et al. “Purposeful sampling for qualitative data collection and analysis in mixed method implementation research”. In: *Adm. Policy Ment. Health Ment. Health Serv. Res.* 42.5 (2015), pp. 533–544.

[25] L. A. Goodman. “Comment: On respondent-driven sampling and snowball sampling in hard-to-reach populations and snowball sampling not in hard-to-reach populations”. In: *Sociol. Methodol.* 41.1 (2011), pp. 347–353.

[30] Hussein A. Alsaeedi and Ahmed S. Alhegami. “An Incremental Interesting Maximal Frequent Itemset Mining Based on FP-Growth Algorithm”. In: *Complexity* 2022.1 (2022), p. 1942517.