**The Paradox of Data Quality and Data Science: When Accuracy Clashes with Practicality**  
**Topic: Data Cleaning vs. Data Integrity – Examining Cases Where Aggressive Cleaning Removes Valuable Outliers or Introduces Bias**

**Authors: Karolina Weyna, Rahul Tak, Ștefan Găman**

**Introduction**

In the data-driven age, the quest for "clean" data is widely regarded as foundational to effective data science. However, the underlying assumption that *cleaner data equates to better insights* is increasingly being challenged by cases where aggressive data cleaning has paradoxically degraded data integrity. The tension between data cleaning and data integrity represents a critical paradox in modern data science, where the goal of accuracy often clashes with practical decision-making. In the upcoming sections we will explore the implications of extensive data cleaning practices, particularly focusing on how the removal of outliers or imposition of uniformity can erase valuable variability, obscure real-world complexities, and even introduce systemic bias into models.

**Understanding the Trade-Offs in Data Cleaning**

Data cleaning typically involves identifying and rectifying errors, missing values, duplicates, and outliers (Rahm & Do, 2000). While these steps are essential in eliminating noise and making datasets analysis-ready, they are rarely value-neutral. Decisions made during cleaning often reflect subjective judgments about what constitutes an error or an anomaly. When datasets are stripped of outliers, for example, this can lead to clean but misleading representations of reality.

A key problem arises when cleaning practices are applied uniformly across diverse data contexts. For example, an outlier in medical data may represent an anomaly—or it might signify a critical indicator of a rare disease. Disregarding such data can undermine the model's ability to detect edge cases, which, in medical or financial domains, can be ethically and operationally catastrophic.

**Case Study 1: Credit Scoring and the Erasure of Marginalized Behavior Patterns**

In a review of credit scoring algorithms used by a major U.S. financial institution, it was discovered that aggressive preprocessing steps such as excluding applicants with irregular income patterns or atypical banking behaviour had disproportionately eliminated low-income and freelance workers from the model’s training data (O’Neil, 2016). These profiles were flagged as “outliers” and excluded from the dataset due to their deviation from conventional financial behaviour.

The result was a model optimised for middle and upper income earners with regular pay cycles, thereby systematically disadvantaging gig economy workers and immigrants. Here, the data cleaning process eliminated noise only from the perspective of the majority population, failing to recognise that what was removed were actually critical signals from underrepresented groups. This type of unintentional bias has implications for fairness and ethical accountability in machine learning (Barocas et al., 2019).

**Case Study 2: Health Informatics and the Loss of Rare Disease Data**

In clinical trials and hospital datasets, outliers often signify complex or rare conditions. One 2018 study involving an AI tool for predicting cardiac events showed that removing high-leverage outliers from EHRs (electronic health records) reduced the tool’s accuracy in detecting rare but fatal arrhythmias (Kwak et al., 2018). The cleaning algorithm labeled extreme vitals as noise or machine error and discarded them.

While the cleaning procedure improved overall model performance metrics such as mean squared error, it did so at the expense of rare-case detection. This reveals a deeper conflict: a cleaner dataset produced a statistically better model, but a clinically worse tool. In health data science, maintaining data integrity may sometimes require preserving anomalies and working with noisy data rather than simplifying it.

**Case Study 3: HSBC – Data Cleaning and the Blind Spot in Money Laundering Detection**

Between 2006 and 2010, HSBC—operating heavily across Asia—was found to have facilitated over $880 million in money laundering for drug cartels, including Mexico’s Sinaloa cartel. A U.S. Senate investigation revealed that HSBC’s internal data pipelines aggressively cleaned and standardized transaction data, stripping out “inconsistent” fields such as country of origin, routing anomalies, and account metadata.

These steps, aimed at simplifying cross-border reporting, also erased key indicators of suspicious activity. As a result, high-risk transactions passed through undetected. Internal alerts were ignored or overwritten by automated systems that favored efficiency over scrutiny.

In 2012, HSBC paid $1.9 billion in fines and entered a deferred prosecution agreement, marking one of the largest AML enforcement actions in history.

Therefore, over-cleaning for speed and uniformity can introduce dangerous blind spots, especially in financial crime detection, where outliers are often red flags—not noise.

**The Bias Introduced by Imputation and Normalization**

Missing data is another domain where data cleaning often introduces hidden bias. Common techniques such as mean imputation or regression imputation are used to fill in gaps. However, these strategies can introduce structural bias by assuming homogeneity in the underlying data distribution. For instance, in demographic datasets, imputing missing values with group averages can flatten socio-economic variance, thereby masking inequalities or distorting policy models (Little & Rubin, 2019).

Similarly, normalisation and standardisation techniques are useful for algorithmic performance but often downplay the meaningful scale of variation. In sentiment analysis, extreme values may reflect intense dissatisfaction or enthusiasm. Truncating these values to conform with standard ranges can reduce the sensitivity of models designed to detect crises or emerging trends.

**Practicality vs. Integrity: A False Binary?**

The recurring theme in the above mentioned cases is the false separation between data accuracy and practicality. Often, decisions are framed as a choice between a clean dataset that’s easy to work with and a messy dataset that’s hard to model. But cuurrent best practices suggest that data cleaning and integrity can coexist provided that transparency and domain knowledge guide the process.

Techniques such as *robust statistics* (Huber, 2011) and *explainable AI* allow for models that can handle noisy or outlier-rich data without compromising fairness or performance. Similarly, leveraging unsupervised anomaly detection, instead of outright removal, can preserve data variability while still flagging potential issues for human review.

**Conclusion**

While data cleaning remains an inherent part of the data science pipeline, it must be applied with caution, especially when decisions risk erasing data that carry meaningful variance. As demonstrated in financial and health domains, aggressive cleaning can introduce bias, reduce inclusivity, and harm the reliability of downstream decision-making. The key lies not in avoiding data cleaning, but in making it context-aware, transparent, and guided by both statistical rigour and ethical foresight. In the paradox of data quality, integrity should not be sacrificed for neatness. Rather, our analytical tools must evolve to embrace the structure of real-world data.

**References**

Barocas, S., Hardt, M., & Narayanan, A. (2019). *Fairness and Machine Learning*. fairmlbook.org.

Huber, P. J. (2011). *Robust Statistics* (2nd ed.). Wiley.

Kwon, J. M., Lee, Y., Lee, Y., Lee, S., & Park, J. (2018). An algorithm based on deep learning for predicting in‐hospital cardiac arrest. Journal of the American Heart Association, 7(13), e008678.

Little, R. J. A., & Rubin, D. B. (2019). *Statistical Analysis with Missing Data* (3rd ed.). Wiley.

O’Neil, C. (2016). *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. Crown Publishing.

Rahm, E., & Do, H. H. (2000). Data Cleaning: Problems and Current Approaches. *IEEE Data Engineering Bulletin, 23*(4), 3–13.

U.S. Senate Subcommittee Report (2012) – *HSBC Case History on Money Laundering.*