Human-Centric Data Cleaning Summary:

The paper establishes the need for keeping Humans involved in the Data Cleaning Process and proposes a vision of how to help humans with Data cleaning. As modern businesses and industries collect large volumes of data, it is extremely important that the data is clean and accurate to help decision-makers. Common Data errors include incorrect entry by humans, duplicates, integrity constraint violation, missing values, etc.

There are many data cleaning systems that involve humans in the cleaning process such as Poter's Wheel, GDR, KATARA, etc. The authors made the following observations on existing data cleaning techniques that, Humans are considered the ultimate authority to verify data validity. Automated tools handle different parts of the data differently. For some sensitive data fields such as Salary, it is preferred that humans make the correction. Existing techniques do not assess the factors that produce data repairs, they also assure that human experts are perfect while in reality there is potential for human error in different parts of the data cleaning process.

The proposed vision for data cleaning should support the following features:

1. Heterogeneity: Different cleaning agents for different data.
2. Isolation: Humans shouldn't be aware of internal cleaning logic
3. Accountability: The system should assess the reliability of different factors involved.
4. Human cost Optimization: Take user cost and expertise into consideration.

The authors presented an architecture for their envisioned system that includes, Four main components: Detectors, Repairers, Cleaning Resources, and Validators. All detectors and repairers should be pluggable black boxes that are applied to raw data views(a subset of raw data). They use cleaning resources such as rules metadata to detect and repair raw data. Cleaning resources (specifications) are supplied by domain experts. Finally, Humans validated the automated repairs. Human feedback is then used by the system to assess its repairs.

They characterized the Human Expertise required for the cleaning process into different roles.

- In Detection Task, Non-Technical user reporting error is Data User.
- In Repair Task, Person with necessary technical skills for repair is Data Curator
- In the Validation task, Humans knowledgeable about the data but are not technical are called Data Validators.
- In the specification Task, The person writing the specification to capture errors needs to be an expert and thus is called Domain Expert.

The paper also proposes a method to measure Human Expertise.

Expertise(h, C, T) = #correct(C, T) / #validated(C, T)

A ratio of correct validated repairs over total validated repairs.

According to the authors, A good interaction model is necessary to optimize the cleaning effort, that should:

1. Minimize communication between human roles.
2. Account for all human-to-human interaction.

The paper then discusses Cross-Agent Cost Optimization, the Consequences of involving human or automatic cleaning agents on Data quality, and how to schedule jobs by either Quantitative Cost Optimization or Qualitative Cost Optimization. The paper then identified different bottlenecks in the different parts of the cleaning process. The presented a scoring measure given the factors.

Quality(f) = #correct(f) / #validated(f)

The ratio of correct validated cells over total validated cells

Related Data cleaning systems were discussed, including NADEEF and KATARA.

The paper was concluded with the question of maintaining privacy in the data cleaning process