

## DATA QUALITY MINING RESEARCH: NEW DIRECTIONS



Laure Berti-Equille



Tamraparni Dasu

**LAURE BERTI-ÉQUILLE** (University of Rennes 1, France)

**TAMRAPARNI DASU** (AT&T Labs – Research, NJ, USA)

### **Abstract:**

Data quality mining is an indispensable part of a data mining task, where data mining methods are used to detect, explain and clean the glitches in data to ensure the reliability of the results of data mining. Our tutorial provides an overview of the state of the art in traditional data quality approaches, and addresses the issue of complex, interdependent glitches within a systematic data quality mining framework. Such an approach is novel and allows us to identify and highlight new research directions in data quality mining.

As data types and data structures change to keep up with evolving technologies and applications, data quality problems too have evolved and become more complex. Data streams, web logs, wikipedias, biomedical applications, video streams and social networking websites generate a mind boggling variety of data types. Data quality mining, the use of data mining to manage, measure and improve data quality, has focused mostly on addressing each category of data glitch separately as a static entity. In this tutorial we highlight: (a) the applicability and effectiveness of the methodologies for various data types such as structured, semi-structured and stream data, (b) the detection of concomitant data glitches, and (c) a general framework for the iterative detection, explanation and cleaning of concomitant multivariate data glitches. We survey past work, introduce current research and tools, and highlight new directions and open research problems in data quality mining. The tutorial includes an elaborate and detailed

real-life case study, numerous applications and practical examples, and an extensive set of references.