From Extensive Data Collection to Intensive Quality Knowledge Production

Leon Reznik, Igor Khokhlov
Department of Computer Science,
Rochester Institute of Technology
E-mail: lr@cs.rit.edu, ixk8996@rit.edu

Problems. Internet measurements build up colossal data collections. This data may have various quality characteristics. Unfortunately, these characteristics are commonly not provided along with the data or ignored. If the current trend continues, we may expect an unprecedented scale of generating, storing, and communicating more and more low-quality data overfilling the available capacities. Without significant changes, existing infrastructure will not scale up these features to massive data arrays, which will have to be communicated, computed, and controlled.

Roads to solutions.

Point #1: *Metadata* that describes the measurement conditions, including sensors, their characteristics, needs to be provided along with the measurement results.

Unfortunately, data integrity could be violated due to hacking attacks.

Point #2: Infrastructure *security* characteristics need to be also provided.

The metadata could grow up to huge volumes.

Point #3: Metadata needs to be processed, producing limited, preferably *numerical metrics* sets, which are given to a data user.

However, the user can still be overwhelmed with these data.

Point #4: Users prefer getting *knowledge* and knowledge-based *service* that help them solve their problems.

To produce this knowledge, data + metadata have to be processed together (see Fig. 1) to get data quality that is, by definition, the degree how much the data fits the user needs.

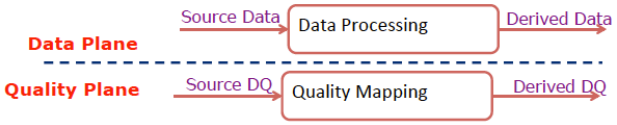

Figure 1. Novel multidimensional data processing and communication

The attention needs to shift from getting more measurements and data to planning and conducting effective and efficient research experiments to produce more knowledge. The progress and innovation may no longer be hindered by the ability to collect data but by the ability to manage, analyze, summarize, visualize, and discover knowledge from the collected data in a timely manner and in a secure and scalable fashion. *Novel data management and knowledge generation principles that should involve processing and filtering the data based on their quality characteristics need to be developed and employed. They have to be supported by network services and protocols facilitating delivery not data only but also data quality (DQ) characteristics.*

Our major solution is to develop methodologies and technologies that will provide an end user or an application with measurement data of a specified quality at the point of use. This goal could be achieved by the dynamic selection, preferably in real-time, of the data sources and communication paths from them to the points of data use. In our NSF funded research (award ACI-1547301), we are building a proof-of-the-concept design, which will be used to develop, verify and promote a comprehensive methodology for DQ evaluation focusing on an integration of cybersecurity with other diverse metrics reflecting DQ, such as accuracy, reliability, timeliness, and security into a single methodological and technological framework. The brief framework operation is presented in fig. 2.

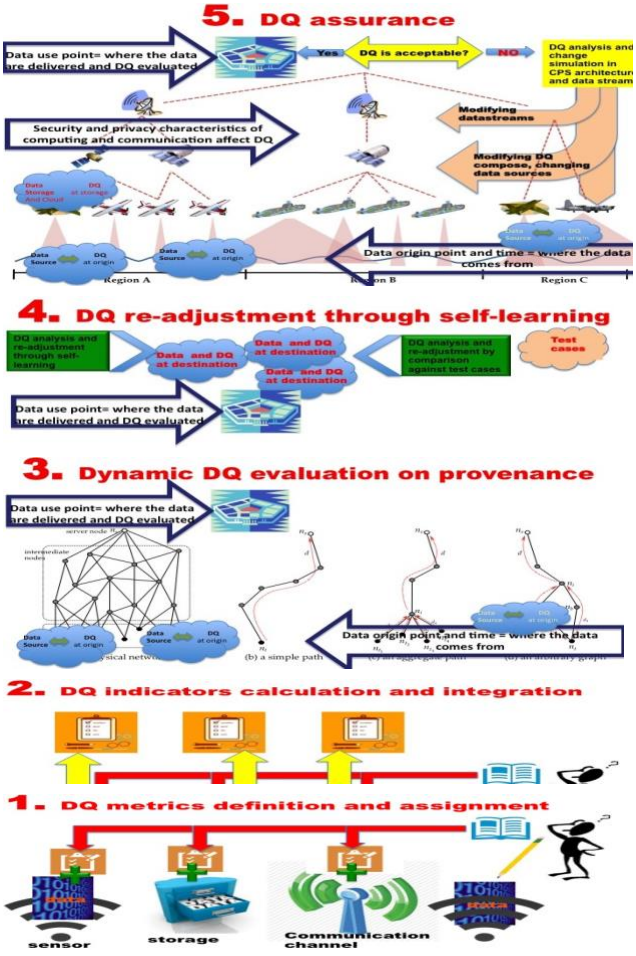

Figure 2. DQ evaluation and assurance framework operation: 1. DQ metrics composition and assignment, 2. DQ initial indicators calculation based on the metrics and their integration, 3. Dynamic DQ changes based on data provenance, 4. DQ integral indicators re-adjustment and 5. DQ assurance procedures