Data Quality for Analytics and Machine Learning (ML)

Wo Chang
Digital Data Advisor
ISO/IEC JTC 1/SC 42/WG 2 Data Working Group, Convenor
National Institute of Standards and Technology, US
wchang@nist.gov

May 24, 2022
Motivations – Why data quality is important?

The success of AI machine learning has led to substantial improvements comparable to human performance and beyond across a variety of application domains. However, data quality for training, validation and verification plays a vital role, especially in the areas of \textit{accuracy, precision, credibility, consistency, completeness, safety, security, privacy/societal/ethical concerns, etc.}

**Sample Domains**

**Healthcare**
- Electronic Health Records
- Precision Medicines
- Drug Discoveries
- Care Management
- Lab Testing Services
- Others...

**Autonomous Vehicles**
- Drones
- Self-driving Cars
- Vehicle-to-Vehicle Communications
- Unmanned Aerial Vehicle
- Others...

**Financial**
- Stock Market
- Electronic Commerce
- Online Transaction
- Banking
- Insurance
- Others...
Data Empower Analytics/AI to Meet Applications Need

Data as oil to analytic/AI engine to solve complex application problems.
Data Quality for Analytics and Machine Learning: Issues

Data (structure, semantic, and metadata) quality impacts (wrong data produce wrong trained models) analytic/AI models results influence applications.
Data Quality for Analytics and Machine Learning: Overall Goal

To provide effective data quality for analytic and machine learning in the areas of measures, process, and management at each stage of the data life cycle across different levels of the organization.

Sample Existing Data Life Cycles Management with Specific Goals

- CRISP-DM (Cross Industry Standard Process for Data Mining)
- SEMMA (Sample, Explore, Modify, Model, and Assess)
- KDD (Knowledge Discovery in Databases)
- DSL (Data Science Lifecycle)
Data Quality for Analytics and Machine Learning: Data Life Cycle

ISO/IEC 8183 provides an overarching data life cycle framework that is instantiable for any AI system that is applicable across different levels of the organization from idea conception to system decommissioning stages with common terminologies and processes.

ISO/IEC 8183 AI Data life cycle framework

1. Idea conception
2. Business requirements
3. Data planning
4. Data acquisition
5. Data preparation
6. Building a model
7. System deployment
8. System operation
9. Data decommissioning
10. System decommissioning

Boundary of data processing

Continuous validation and verification of system

Validation and verification of model

ISO/IEC 8183 AI Data life cycle framework (under DIS balloting)
ISO/IEC 5259-x Data Quality for Analytics and Machine Learning

A holistic approach is needed to oversee the implementation and operation of data quality measures, data quality management requirements and guidelines, and data quality process for various types of analytics and machine learnings with adequate controls throughout the ISO/IEC 8183 AI Data Life Cycle Framework.

Leverage Available Standards

- ISO/IEC 8183*
- ISO/IEC 25012
- ISO/IEC 25024
- ISO/IEC 38500
- ISO/IEC 38502
- ISO/IEC 38505
- ISO/IEC 38507
- ISO/IEC 8000-61

ISO/IEC 23053
ISO/IEC 22989
ISO/IEC 25024
ISO/IEC 38502
ISO/IEC 38505
ISO/IEC 38507
ISO/IEC 8000-61

ISO/IEC 23053
ISO/IEC 22989
ISO/IEC 25024
ISO/IEC 38502
ISO/IEC 38505
ISO/IEC 38507
ISO/IEC 8000-61

* – under DIS balloting

ISO/IEC 8183 AI Data life cycle framework (under balloting)
ISO/IEC 5259-1 Overview, terminology, and examples

**Scope:** This document provides the means for understanding and associating the individual documents of the ISO/IEC “Artificial intelligence — Data quality for analytics and ML” series and is the foundation for conceptual understanding of data quality for analytics and machine learning. It also discusses associated technologies and examples (e.g. use cases and usage scenarios).

Key

- Stage where data are processed
- Iteration
- Primary development pathway
- Feedback pathway

**Data life cycle framework (from ISO/IEC 8183)**

**Instantiations of DLC model for 5259 series**
## ISO/IEC 5259-1 Overview, terminology, and example

### DQ Concept Framework for Analytics and Machine Learning

<table>
<thead>
<tr>
<th>DQ Concept</th>
<th>ISO/IEC Standards</th>
</tr>
</thead>
<tbody>
<tr>
<td>DQ model</td>
<td>ISO/IEC 5259-1</td>
</tr>
<tr>
<td>DQ measures</td>
<td>ISO/IEC 5259-2</td>
</tr>
<tr>
<td>DQ assessment</td>
<td>ISO/IEC 5259-3</td>
</tr>
<tr>
<td>DQ improvement</td>
<td>ISO/IEC 5259-4</td>
</tr>
</tbody>
</table>

- **DQ model**
- **DQ measures**
- **DQ assessment**
- **DQ improvement**

**DQ management**

**DQ reporting**
- ISO/IEC 5259-2
- ISO/IEC 5259-4

**DQ governance**

**Data provenance**
- ISO 8000-120
ISO/IEC 5259-2 Data Quality Measures

Scope: This document provides a data quality model, data quality measures and guidance on reporting data quality in the context of analytics and machine learning (ML). This document builds on ISO 8000 series, ISO/IEC 25012:2008 and ISO/IEC 25024. The aim of this document is to enable organizations to achieve their data quality objectives and is applicable to all types of organizations.
ISO/IEC 5259-2 Data Quality Measures

Data Quality Measurement Framework
(under development)

DQ Characteristics
1. Accuracy
2. Precision
3. Completeness
4. Representativeness
5. Consistency
6. Relevance
7. Data scalability
8. Context coverage
9. Portability
10. Timeliness
11. Currentness
12. Identifiability
13. Auditability
14. Credibility
15. Understandability
16. Balance
17. Effectiveness
18. Similarity

ISO/IEC 20512
ISO/IEC 20522, 23, 24
ISO/IEC 20521
ISO/IEC 5259-3 DQ Management Requirements & Guidelines

Scope: This document specifies requirements and provides guidance for establishing, implementing, maintaining and continually improving the quality for data used in the areas of analytics and machine learning. This document does not define a detailed process, methods or metrics. Rather it defines the requirements and guidance for a quality management process along with a reference process and methods that can be tailored to meet the requirements in this document. The requirements and recommendations set out in this document are generic and are intended to be applicable to all organizations, regardless of type, size or nature.
ISO/IEC 5259-4 Data Quality Process Framework

Scope: This document provides general common organizational approaches, regardless of type, size or nature of the applying organization, to ensure data quality for training and evaluation in analytics and machine learning. It includes guidelines for (a) supervised machine learning with regard to the labelling of data used for training machine learning systems, including common organizational approaches for training data labelling; (b) unsupervised machine learning; (c) semi-supervised machine learning; and (d) reinforcement machine learning. This document is applicable to training and evaluation data that comes from different sources, including data acquisition and data composition, data pre-processing, data labelling, evaluation, and data use. This document does not define specific services, platforms or tools.
ISO/IEC 5259-4 Data Quality Process Framework

Typical data labelling methods can include object annotation, bounding box, key-point annotation, instance segmentation and semantic segmentation. Data can be labelled manually by a human annotator or automatically by using pseudo-labelling. Pseudo-labelling is the process of using a model trained on labelled data to predict labels for unlabeled data.
ISO/IEC 5259-5 Data Quality Governance Framework

Scope: This document provides a data quality governance framework for analytics and machine learning to enable governing bodies of organizations to direct and oversee the implementation and operation of data quality measures, management, and related processes with adequate controls throughout the data life cycle. This document can be applied to any analytics and machine learning. This document does not define specific management requirements or process requirements specified in 5259-3 and 5259-4 respectively.
ISO/IEC 5259-5 Data Quality Governance Framework

Relationship between Data Quality Governance Framework and 5259 projects

ISO/IEC AI Workshop Series, Novel AI Standardization Approaches Track, Wo Chang, NIST/ITL, May 24, 2022
ISO/IEC 5259-6 Data Quality Visualization

**Scope:** This document is an overview of data visualization within the context of Artificial Intelligence (AI) and Machine Learning (ML) applications. It is intended to provide examples of where data visualization may be employed by various stakeholders at different stages of the AI life cycle.

Concept of Framework for Data-Quality-Driven Visualization for AI

![Diagram of Concept of Framework for Data-Quality-Driven Visualization for AI](Source: ISO/IEC 20521:2012)

Figure 5 — Relationship among property to quantify, measurement method, QME and QM
(Source: ISO/IEC 20521:2012)
### ISO/IEC 5259-6 Data Quality Visualization

<table>
<thead>
<tr>
<th>Sample Data Quality Measure from ISO/IEC 25012</th>
<th>Inherent/System Dependent</th>
<th>Phase in System Life Cycle</th>
<th>Stakeholders</th>
<th>Quality Measurement Requirements</th>
<th>Visualization Requirements</th>
<th>Applicable Visualization Methods</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
<td>I</td>
<td>Inception, Development</td>
<td>Developer, Consumer</td>
<td>Sampling reflects the real world</td>
<td>Index or bar chart</td>
<td>Chart/graph</td>
<td>Multiplot of features</td>
</tr>
<tr>
<td>Completeness</td>
<td>I</td>
<td>Inception</td>
<td>Developer, Regulator</td>
<td>Statistically large data sample</td>
<td>Index</td>
<td>Chart/graph</td>
<td>Multiplot of features</td>
</tr>
<tr>
<td>Consistency</td>
<td>I</td>
<td>Inception, Development, Deployment</td>
<td>Developer</td>
<td>Data samples have same dimensions and features</td>
<td>Index/bar chart</td>
<td>Chart/graph</td>
<td>Group cluster plots</td>
</tr>
<tr>
<td>Credibility</td>
<td>I</td>
<td>Deployment</td>
<td>Consumer</td>
<td>Limited bias and impartial sampling</td>
<td>Index</td>
<td>Chart</td>
<td>Histogram</td>
</tr>
<tr>
<td>Current-ness</td>
<td>I</td>
<td>Inception thru Deployment</td>
<td>Developer, Regulator</td>
<td>Age of the data remains reflective of reality/use case</td>
<td>Time date</td>
<td>Numerical</td>
<td>Numerical</td>
</tr>
<tr>
<td>Accessibility</td>
<td>I/D</td>
<td>Development</td>
<td>Regulator, Consumer</td>
<td>High up time and available to stakeholder</td>
<td>N/A</td>
<td>N/A</td>
<td>Index by access method and stake holder/privilege</td>
</tr>
<tr>
<td>Compliance</td>
<td>I/D</td>
<td>All</td>
<td>Regulator, Developer</td>
<td>Regulator ensures data gathered per applicable norms</td>
<td>N/A</td>
<td>Index chart of criteria</td>
<td>Glyph of compliance criteria</td>
</tr>
<tr>
<td>Confidentiality</td>
<td>I/D</td>
<td>All</td>
<td>Consumer, Regulator, Developer</td>
<td>Domain dependent</td>
<td>N/A</td>
<td>Numerical/index</td>
<td>N/A</td>
</tr>
<tr>
<td>Efficiency</td>
<td>I/D</td>
<td>All</td>
<td>Developer</td>
<td>Only what is used to answer a question/problem is in the data</td>
<td>N/A</td>
<td>Index</td>
<td>N/A</td>
</tr>
<tr>
<td>Precision</td>
<td>I/D</td>
<td>Development</td>
<td>Developer</td>
<td>Meets the intended application</td>
<td>Numerical</td>
<td>Numerical</td>
<td>N/A</td>
</tr>
<tr>
<td>Traceability</td>
<td>I/D</td>
<td>All</td>
<td>Regulator, Developer</td>
<td>Lineage is tracked</td>
<td>Index</td>
<td>Chart</td>
<td>Flow chart</td>
</tr>
<tr>
<td>Understandability</td>
<td>I/D</td>
<td>All</td>
<td>Regulator, Consumer</td>
<td>A stake holder can assess the data</td>
<td>Index or binary</td>
<td>Index</td>
<td>Multiplot by stake holder</td>
</tr>
<tr>
<td>Availability</td>
<td>D</td>
<td>All</td>
<td>Consumer</td>
<td>Data exists</td>
<td>Index</td>
<td>Index</td>
<td>Check list</td>
</tr>
<tr>
<td>Portability</td>
<td>D</td>
<td>Development</td>
<td>Developer</td>
<td>Data can be transferred to other systems for other application</td>
<td>Check list</td>
<td>Check list</td>
<td>N/A</td>
</tr>
<tr>
<td>Recoverability</td>
<td>D</td>
<td>Development</td>
<td>Producer, Developer</td>
<td>Data is stored in a manner as to be redundant and errors can be fixed</td>
<td>Check list</td>
<td>Check list</td>
<td>N/A</td>
</tr>
</tbody>
</table>
## ISO/IEC 8183 and ISO/IEC 5259-x Projects Timeline

<table>
<thead>
<tr>
<th>ISO/IEC Standard</th>
<th>Approved</th>
<th>DTR/WD</th>
<th>TR/CD</th>
<th>DIS</th>
<th>FDIS</th>
<th>IS</th>
</tr>
</thead>
<tbody>
<tr>
<td>8183 (IS) AI data life cycle framework</td>
<td>02/07/22</td>
<td>--</td>
<td>02/07/22</td>
<td>05/01/22</td>
<td>12/01/22</td>
<td>05/01/23</td>
</tr>
<tr>
<td>5259-1 (IS) Data quality for analytics and machine learning (ML): Part 1: Overview, terminology, and examples</td>
<td>07/30/20</td>
<td>12/21/20 WD23</td>
<td>08/21/22</td>
<td>03/21/23</td>
<td>01/28/24</td>
<td>05/27/24</td>
</tr>
<tr>
<td>5259-2 (IS) Data quality for analytics and machine learning (ML): Part 2: Data quality measures</td>
<td>12/31/20</td>
<td>06/17/21 WD14</td>
<td>02/13/23</td>
<td>09/20/23</td>
<td>04/20/24</td>
<td>09/20/24</td>
</tr>
<tr>
<td>5259-3 (IS) Data quality for analytics and machine learning (ML): Part 3: Data quality management requirements and guidelines</td>
<td>07/30/20</td>
<td>12/21/20 WD12</td>
<td>08/21/22</td>
<td>03/21/23</td>
<td>01/28/24</td>
<td>05/27/24</td>
</tr>
<tr>
<td>5259-4 (IS) Data quality for analytics and machine learning (ML): Part 4: Data quality process framework</td>
<td>07/30/20</td>
<td>12/21/20 WD14</td>
<td>08/21/22</td>
<td>03/21/23</td>
<td>01/28/24</td>
<td>05/27/24</td>
</tr>
<tr>
<td>5259-5 (IS) Data quality for analytics and machine learning (ML): Part 5: Data quality governance framework</td>
<td>02/17/22</td>
<td>05/31/22 WD1</td>
<td>03/13/23</td>
<td>03/11/24</td>
<td>10/11/24</td>
<td>03/10/25</td>
</tr>
</tbody>
</table>
SC 42/WG 2 Overview and Officers

Title: Data

Terms of Reference: Standardization in relation to data in the context of artificial intelligence, big data, and data analytics

Officers
- Convenor: Wo Chang
- Secretary: Heather Benko
- Editors: Gautam Banerjee (24668), Colin Crone (8183 PWI), Suwook Ha (5259-1), KyoungSook Kim (5259-2), Martin Saerbeck (5259-3), Wanzhong Ma (5259-4), Gyeung-Min Kim (5259-5)

Membership
310+ Experts from 27 National Bodies: Australia, Austria, Belgium, Canada, China, Denmark, Finland, France, Germany, India, Ireland, Israel, Italy, Japan, Korea, Luxembourg, Netherlands, Norway, Poland, Portugal, Russian Federation, Singapore, Spain, Switzerland, United Arab Emirates, United Kingdom, United States
Questions?

Please contact: wchang@nist.gov