# Measuring the 3V's of Big Data: A Rigorous Approach

*Olga Ormandjieva[1], Mandana Omidbakhsh[2] and Sylvie Trudel[2]*

*[1] Concordia University*

*[2] Université du Québec à Montréal*

# Agenda

- Quality of Big Data
- Research Goal
- Background and related work
  - Overview of our Approach to Big Data Quality Measurement
  - Big Data Characteristics Hierarchy: Relevance of NIST
- Proposed  3Vs measurement hierarchical model
- Illustration and validation of the measurement model
- Future work

# QUALITY OF BIG DATA

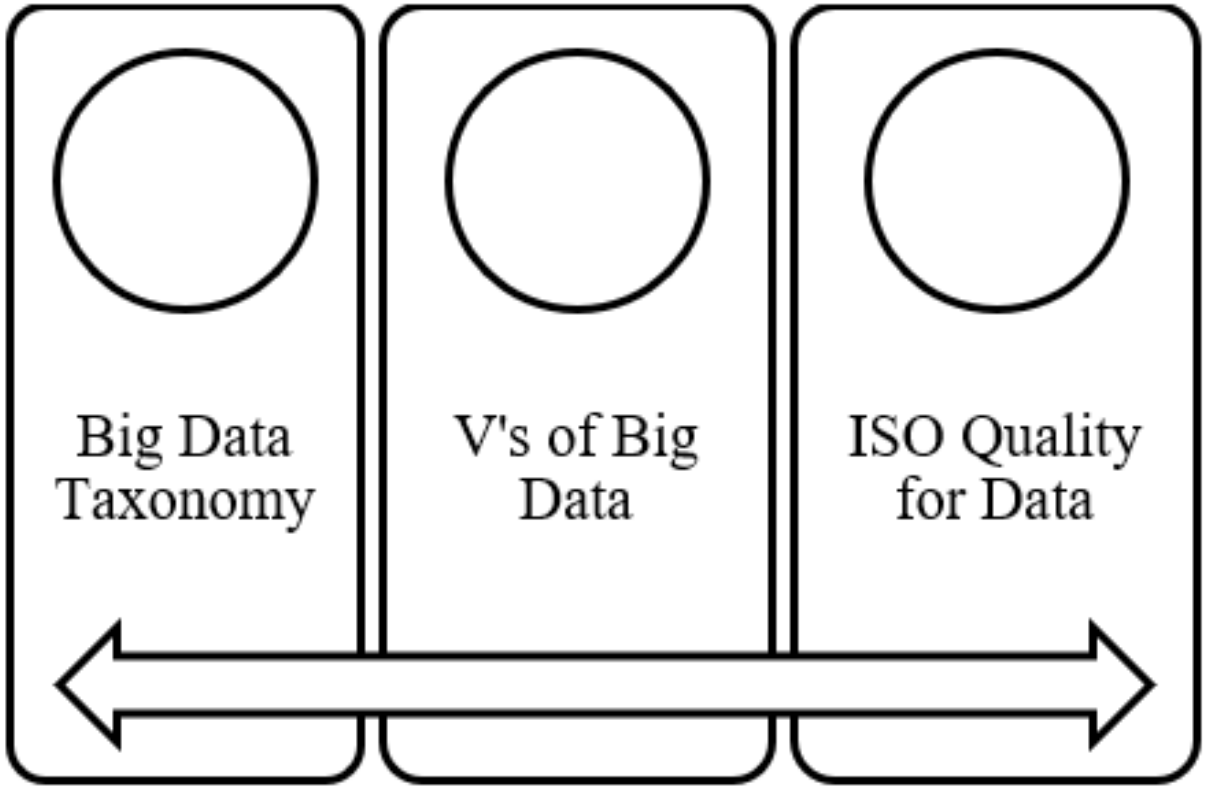
- Data is generally considered as high quality if it is "fit " for its intended uses in operation, decision making and planning
- The quality of Big Data can be measured in terms of its characteristics (the V's)
- V's selected in this work (10V's):
  - Volume, Velocity, Variety, Variability, Value, Veracity, Validity, Vulnerability, Visualisation, Volatility

# Research Goal

- Research Project: ensuring visibility and transparency of big data quality by assessing quantitatively the quality of big data in terms of its industrially recognized characteristics (the V's)
- This paper: Proposing a novel measurement information model for the 3V's
  - Volume: Magnitude of the data
  - Variety: Different forms of data
  - Velocity: Speed at which data is being generated.

# Overview of our Approach to Big Data Quality Measurement

Omidbakhsh, M., Ormandjieva, O.: Toward A New Quality Measurement Model for Big Data, DATA 2020, The 9th International Conference on Data Science, Technology and Applications (2020).



# Relevance of ISO 25012 and ISO25024

- The existing standards reflect two main views:
  - The degree to which data quality characteristics satisfy data requirements
  - The degree to which data  quality is reached and preserved when data is used under specified conditions
  - However, quality measurement of Big data is not addressed

# Relevance of NIST to Big Data Quality Measurement

- NIST provides  a framework which can be used across industry, academia and researchers for which one can extract meaning from ever-larger and more varied datasets

- This framework created by NIST forms the basis for which our research is built and so will be adopted in our research

- The NIST taxonomy for Big Data which consist of Elements, Records, Dataset and Multiple Dataset will be used as foundation for the proposed hierarchical measurement model.

# NIST: Big Data Characteristics Hierarchy

- Data Element : Single value of data (cell in a csv)
- Records: Multiple data elements grouped together (row in a csv)
- Datasets: A collection of records (The entire csv)
- Multiple Datasets: Collection of Datasets (Multiple CSVs)

# Agenda

- Quality of Big Data
- Research Goal
- Background and related work
  - Overview of our Approach to Big Data Quality Measurement
  - Big Data Characteristics Hierarchy: Relevance of NIST
- **Proposed 3Vs measurement hierarchical model**
- Illustration and validation of the measurement model
- Future work

# Mathematical Modeling of the Measurement Entities: set-theoretical approach

- Data Elements: a collection of heterogeneous data elements (set *DE* of $UID_E$)
  - Data elements are stored in records
- record *r* is a 2-tuple *(DEr, m)*
  - *DEr* is a subset of *DE, m* is the *multiplicity*
- Dataset *DS*: a set of records' unique identifiers ($UID_R$)
- Big data is viewed as multiple datasets MDS (in mathematical terms, as a set of multisets).

# Base Measures

- *Number of Distinct Data Elements (Ndde).*
  - Reflects the variety of the dataset
  - Ndde(DE) = |DE|
- *Number of records in a dataset (Nrec)*
  - Assesses the variety of dataset in terms of diversity of records & their sources
  - NREC(DS) = |DS|
- *Number of Datasets in Big Data (NDS)*
  - Number of total unique identifiers $UID_{DST}$ of datasets in multiple datasets
  - Nds(MDS) = |MDS|
  - Assesses the variety of datasets in terms of multiple datasets (MDS)
- *Time (T)*

# Derived Measures and Indicators

- *Length of Big Data (Ldb):* the total number of records in MDS.

$$Lbd\ (MDS) = \sum_{\forall DS \in MDS} Nrec\ (DS)$$

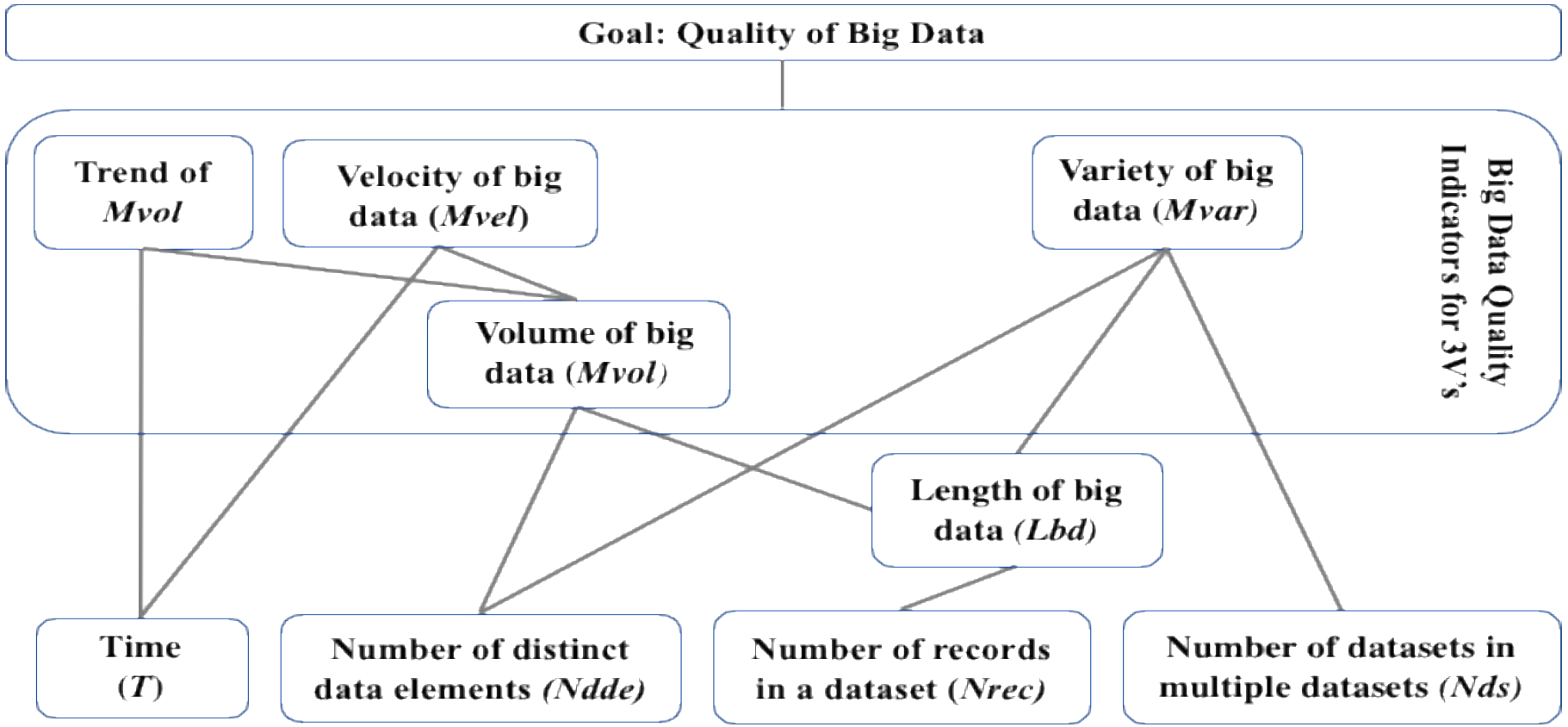
- *Big Data Volume (Mvol):* information content of multiple datasets
  - $Mvol\ (MDS) = Lbd\ (MDS)\ log_2\ (bin\_dec\ (Ndde\ (DE)))$
  - Trend of *Mvol* depicts graphically Volume of big data over time

# Derived Measures and Indicators

- *Big Data Velocity (Mvel):*
  - relative growth of big data over a period of Time *T*
  - $Mvel\ (MDS) = ((Mvol\ (\mathrm{MDS}_{T2}) - Mvol\ (\mathrm{MDS}_{T1})) / Mvol\ (MDS_{T1}) * 100$

- *Big Data Variety (Mvar).*
  - 3-tuple (Ndde (DE), Lbd (MDS), Nds (MDS))

# Hierarchy of the 3V's Measures and Indicators



# Agenda

- Quality of Big Data
- Research Goal
- Background and related work
  - Overview of our Approach to Big Data Quality Measurement
  - Big Data Characteristics Hierarchy: Relevance of NIST
- Proposed 3Vs measurement hierarchical model
- **Illustration and validation of the measurement model**
- Future work

# Measurement Illustration (case 1)

Ndde =35 UIDs

Nrec =20 $UID_R$

Nds = 1 $UID_{DST}$

Lbd= 20 $UID_R$

Mvol ($MDS_{T1}$) =20$log_2$(35) =103 information bits.

| Age | Nationality | Own Pic/week | Exposure | Status/ week | Ratio |
|---|---|---|---|---|---|
| 21 | Saudi | 7 | High | 3 | 0.1 |
| 19 | Pakistan | 2 | High | 12 | 5 |
| 33 | Yemen | 7 | High | 2 | 4 |
| 5 | Saudi | 2 | High | 7 | 0.1 |
| 19 | Pakistan | 2 | Medium | 12 | 3 |
| 25 | Yemen | 3 | Medium | 3 | 4 |
| 22 | Saudi | 7 | Medium | 7 | 0.1 |
| 33 | Pakistan | 8 | Medium | 8 | 5 |
| 21 | Saudi | 7 | Medium | 22 | 5 |
| 27 | Saudi | 8 | Medium | 33 | 1 |
| 22 | Saudi | 2 | Low | 22 | 7 |
| 26 | Saudi | 8 | Low | 2 | 2 |
| 21 | Saudi | 7 | High | 3 | 0.1 |
| 19 | Pakistan | 2 | High | 12 | 5 |
| 33 | Yemen | 7 | High | 2 | 4 |
| 35 | Saudi | 2 | High | 7 | 0.1 |
| 19 | Pakistan | 2 | Medium | 12 | 3 |
| 25 | Yemen | 3 | Medium | 3 | 4 |
| 22 | Saudi | 7 | Medium | 7 | 0.1 |
| 33 | Pakistan | 8 | Medium | 8 | 5 |

# Measurement Illustration (case 2 - Additional Entries to the Sample Data)

Ndde =35 UIDs

$DE_{T1} = DE_{T2}$

Nrec =24 $UID_R$

Nds = 1 $UID_{DST}$

Lbd ($MDS_{T2}$) =24 $UID_R$

Mvol ($MDS_{T2}$) =$24\log_2(35)$ =124 information bits.

Mvel (MDS) = (124 - 103)/ 103 *100 = 20.4%

Mvar (MDS) is (35 data elements, 20 records, 1 dataset).

| Age | Nationality | Own Pic/week | Exposure | Status/ week | Ratio |
|---|---|---|---|---|---|
| 21 | Saudi | 7 | High | 3 | 0.1 |
| 19 | Pakistan | 2 | High | 12 | 5 |
| 33 | Yemen | 7 | High | 2 | 4 |
| 5 | Saudi | 2 | High | 7 | 0.1 |
| 19 | Pakistan | 2 | Medium | 12 | 3 |
| 25 | Yemen | 3 | Medium | 3 | 4 |
| 22 | Saudi | 7 | Medium | 7 | 0.1 |
| 33 | Pakistan | 8 | Medium | 8 | 5 |
| 21 | Saudi | 7 | Medium | 22 | 5 |
| 27 | Saudi | 8 | Medium | 33 | 1 |
| 22 | Saudi | 2 | Low | 22 | 7 |
| 26 | Saudi | 8 | Low | 2 | 2 |
| 21 | Saudi | 7 | High | 3 | 0.1 |
| 19 | Pakistan | 2 | High | 12 | 5 |
| 33 | Yemen | 7 | High | 2 | 4 |
| 35 | Saudi | 2 | High | 7 | 0.1 |
| 19 | Pakistan | 2 | Medium | 12 | 3 |
| 25 | Yemen | 3 | Medium | 3 | 4 |
| 22 | Saudi | 7 | Medium | 7 | 0.1 |
| 33 | Pakistan | 8 | Medium | 8 | 5 |

| | | | | | |
|---|---|---|---|---|---|
| 21 | Saudi | 7 | Medium | 22 | 5 |
| 27 | Saudi | 8 | Medium | 33 | 1 |
| 22 | Saudi | 2 | Low | 22 | 7 |
| 26 | Saudi | 8 | Low | 2 | 2 |

# Validation of 3Vs

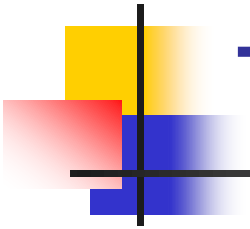
- Measurement validation is "the act or process of ensuring that (a measure) reliably predicts or assesses a quality factor"
- Two approaches to validation have been prescribed and practiced in software engineering: (a) theoretical validation, and (b) empirical validation (not in the scope of this paper)

# Tracking and Consistency

- **Mvol -** the more information big data contains (that is, the higher the information content of the big data), the larger the Mvol indicator value:
  - Mvol ($MDS_{T2}$) > Mvol ($MDS_{T1}$)
- **Mvel -** anticipated increase (20.4%)
- **Mvar -** Mvar ($MDS_{T2}$) is (35 $UID_E$, 24 $UID_R$, 1 $UID_{DST}$), as expected, shows an increase in the MDS length parameter only (was 20 $UID_R$) as compared to Mvar ($MDS_{T1}$)

# Conclusions and Future Work

- A new measurement information model to quantify three aspects of Big Data – Volume, Variety, and Velocity
- Theoretical validation of these 3V's have been demonstrated.
- The model is suitable for big data in any forms of structured, unstructured, and semi-structured
- Empirical validation of these measures with open-access data and industry data.
- The automation of the 3V's measurements



# Thank You!

- Questions?