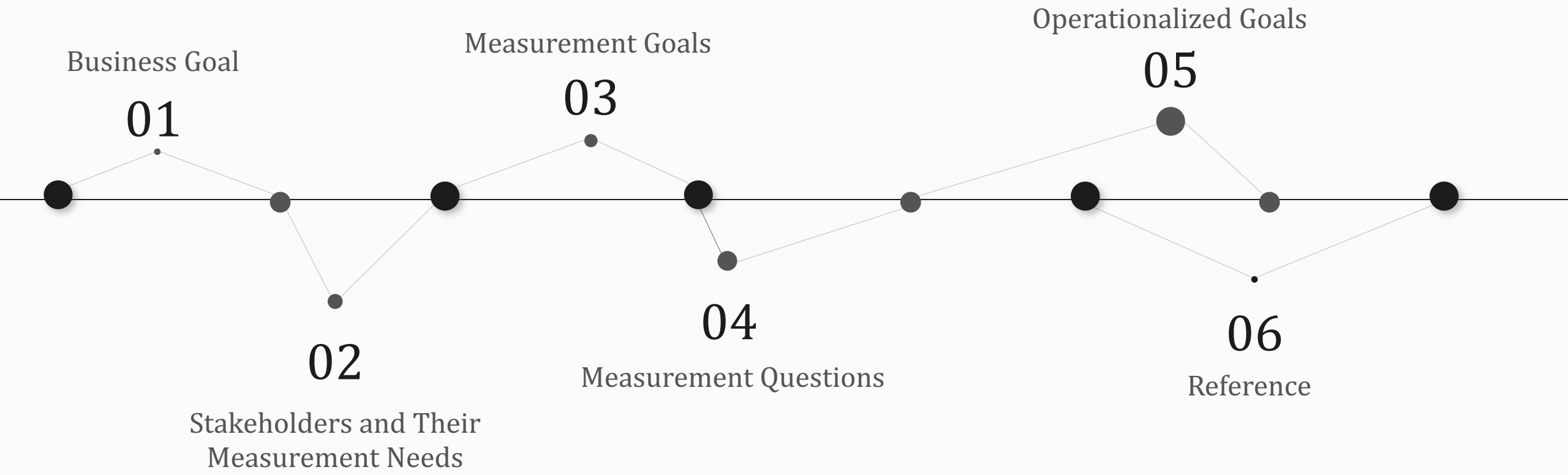
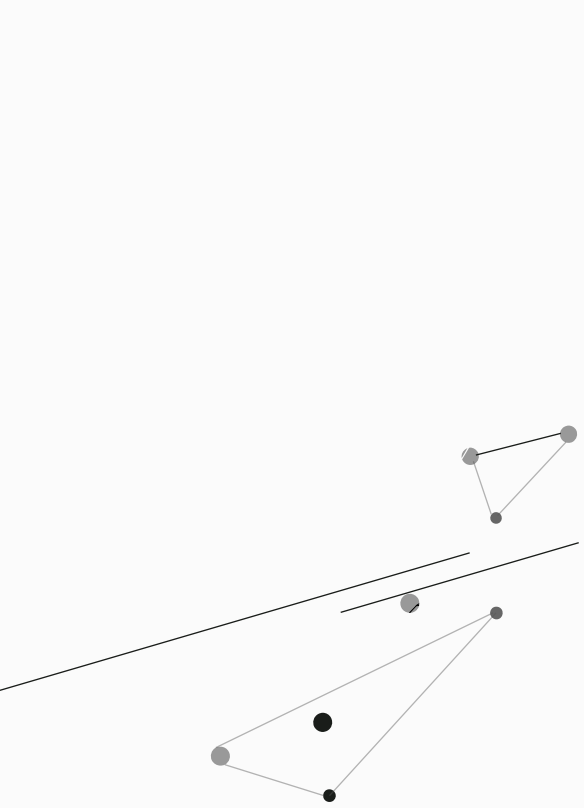
SOEN 6611

Project Step 1 & 2

- Team 5
  Haochen Zou 40158179

Catalogue


Business Goal
01
02
Stakeholders and Their Measurement Needs
Measurement Goals
03
04
Measurement Questions
Operationalized Goals
05
06
Reference



1

Business Goal

- **Six Business Goals** *matched by*
- **Six Measurement Goals** *which contains measures of*
- **Fifteen Big Data Qualities**: Accuracy, Completeness, Consistency, Credibility, Currentness, Accessibility, Compliance, Confidentiality, Efficiency, Precision, Traceability, Understandability, Availability, Portability, Recoverability [1].

01

Business Goal BG #1:

Determine the suitable quantity of data elements is available in datasets for the machine learning pipeline implementation.

(Data Quantity)

Business Goal BG #2

Determine whether the particularity of the big dataset is up to date, appropriate promptly for usage, and complies with current existing relevant regulations.

(Data Timeliness)

Business Goal BG #3:

Determine the structural and format integrity of the big dataset associated with the entity and attribute for the context of the usage.

(Data Organization)

01

Business Goal BG #5

Improve the correctness, comprehensible, and understandable of the dataset.
(Data Correctness)

Business Goal BG #4:

Improve the content trustworthiness, accuracy, reliability, and precision of the big dataset for the machine learning implementation.
(Data Trustworthiness)

Business Goal BG #6:

Improve the connectivity and linkage of the big dataset. Find the common or semantically equivalent or related attributes to link data sources together.
(Data Linkability)

2

Stakeholders and Measurement Needs

- There are **six kinds of positions** who will involve in the project, including algorithm designers, big data managers, big data developers, data analysts, data scientists, and machine learning programmer
- Definitions and detailed information about stakeholders will be discussed in the project report

02

Big data managers

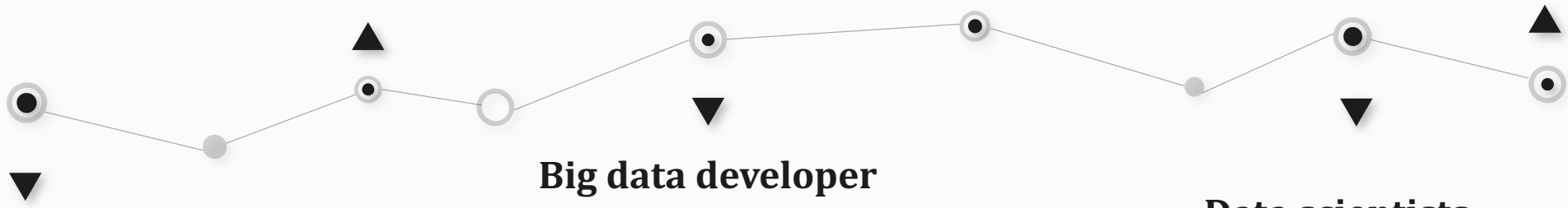
Volume of the datasets and **Vincularity** of the datasets.

Data analysts

Volume, Velocity, and **Variety** of the datasets

Machine learning Programmers

Veracity, Validity, and **Vincularity** so as to make decisions



Algorithm designers

Velocity, because the processing speed is of vital importance to the performance of the machine learning performance.

Variety, since some data elements might not contribute to the algorithm.

Veracity and **Validity** are responsible for the trustworthiness and the accuracy of the extensive dataset respectively, which should also be included.

Big data developer

Veracity of the datasets and **Validity** of the datasets.

Data scientists

The expected benefits from the usage of the measurement results and the summary of the reason for the related efforts might involve **Vincularity**, which is the connectivity and linkage of data. **Vincularity** can benefit them to better structure the datasets. **Variety** should also be included for a better understanding of the datasets.

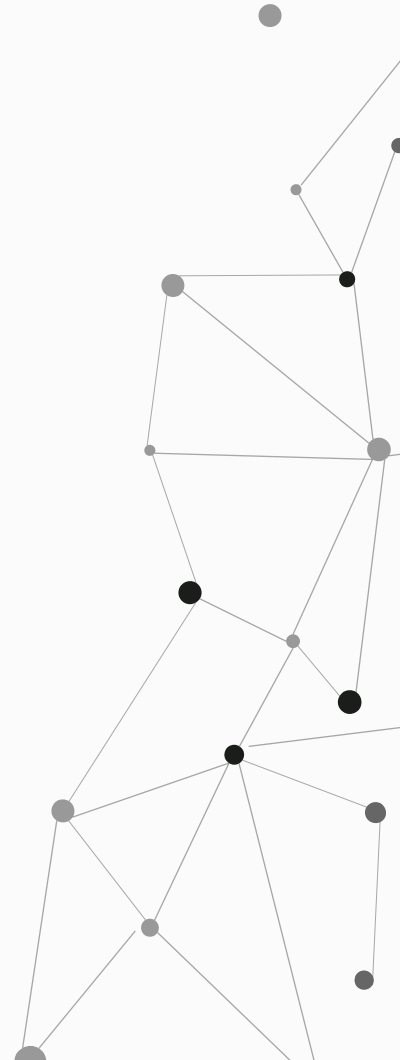
3

Measurement Goals

- Measurement goals are derived from the measurement needs [2].

03

Measurement Goal Label:	Description	Corresponding business goal
MG1 **Big Data Volume**	**Big Data Volume** ***Mvol*** measures the number of information bits across all records required to specify the information content of multiple datasets. Refers to the vast amounts of data that are generated in different time.	Business Goal **BG #1 (Determine Quantity)**
MG2 **Big Data Velocity**	**Big Data Velocity** ***Mvel*** refers to the speed of processing of data in any form of handling, recording, and publishing of data. Velocity also refers to the speed at which data is being generated.	Business Goal **BG #2 (Determine Timeliness)**
MG3 **Big Data Variety**	**Big Data Variety** ***Mvar*** refers to the ever-increasing different forms that data can come in. It reflects correspondingly the diversity of unique data elements, diversity of records, and diversity of datasets.	Business Goal **BG #3 (Improve Organization)**
MG4 **Big Data Veracity**	**Big Data Veracity** ***Mver*** refers to the degree that data is accurate, trusted, and precise. It is not only the accuracy of the data itself but the trustworthiness of the data source, type, and processing of it.	Business Goal **BG #4 (Improve Trustworthy)**
MG5 **Big Data Validity**	**Big Data Validity** ***Mval*** refers to the accuracy of data for the purpose of usage. Validity of data might be having same ideas with veracity of data, but they don't have same concepts and theories.	Business Goal **BG #4 (Improve Correctness)**
MG6 **Big Data Vincularity**	**Big Data Vincularity** ***Mvar*** refers to the connectivity and linkage of data.	Business Goal **BG #5 (Improve Linkability)**

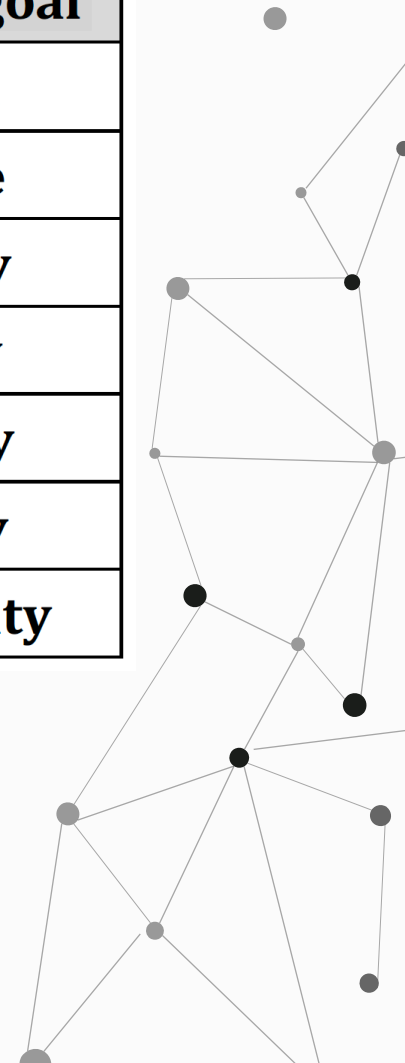


4

Measurement Questions

04

Question Label	Description	Corresponding measurement goal
Q0	What are the multiple datasets for big data analysis?	**Big Data Set**
Q1	What is the number of information bits in different time?	**MG1 Volume**
Q2	What is the speed data being processed and generated?	**MG2 Velocity**
Q3	What is the diversity of components in the dataset?	**MG3 Variety**
Q4	What is the trustworthiness of the data elements?	**MG4 Veracity**
Q5	What is the accuracy of the data elements?	**MG5 Validity**
Q6	What is the linkage among data elements in datasets?	**MG6 Vincularity**

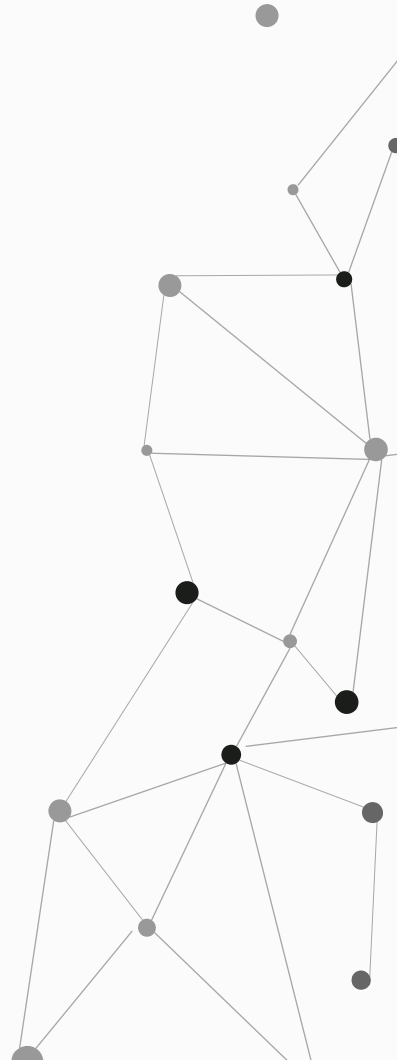


5

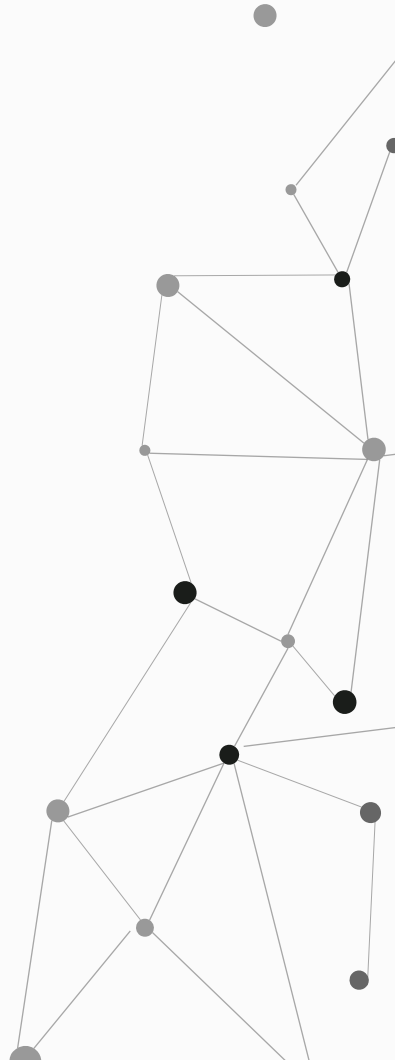
Operationalized Goals

05

Operationalized Goal: Label and Description	**OG #1**: Determine the amount of data available among all records required to specify the content of multiple datasets and have it ready for the machine learning pipeline implementation.
Corresponding Measurement Goal Label	**MG #1**: Big Data Volume.
Object of Interest	The big data information product and resource.
Purpose	Big data volume refers to the magnitude of data. The process of this operationalized goal is to analyze the quantity of information bits around entire records required in order to specify information content in datasets. The purpose of this operationalized goal is to compare and comprehend multiple datasets in terms of their information content, as well as to oversight their growth over time [1].
Quality Focus, Perspective	Quality focus on examining the data scale and quality. The perspective is from the viewpoint of big data managers and data analysts are supposed to analyze the general trend of the amount of data at different time spans during the implementation.
Environment and Constraints	The big data program and the machine learning pipeline. The factors and parameters that should take into consideration are application factors, resource factors, and process factors.

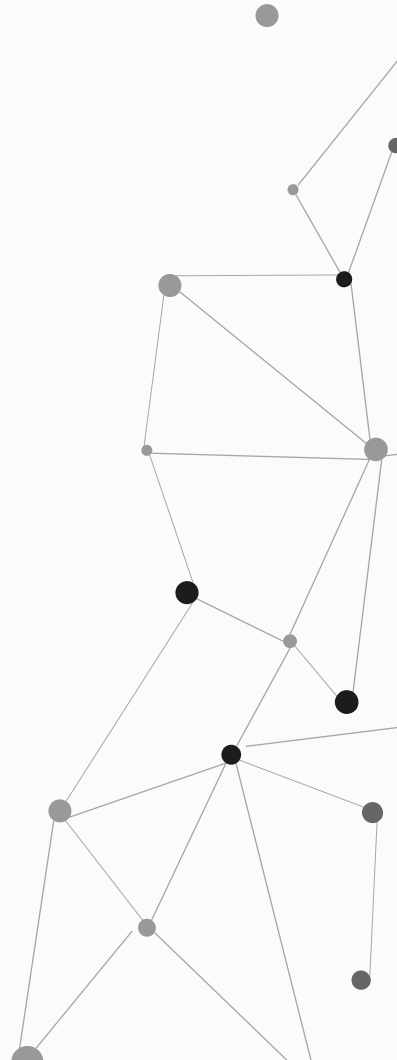


Operationalized Goal: Label and Description	**OG #2**: Dynamic aspect of the big dataset [2]. Determine the speed of real-time processing of data in order to ensure that the data is timeliness and not out-of-date in the further usage.
Corresponding Measurement Goal Label	**MG #2**: Big Data Velocity.
Object of Interest	The big data information product and resource.
Purpose	Big data velocity refers to the emphasis on the speed of data streams, data processing, data generations and so on. The purpose of this process is to evaluate the data in motion before being prepared for retrieval and characterize the speed of receiving data in order to control the data loss as well as understand the lifetime span utility of the data at various time frames in the implementation.
Quality Focus, Perspective	The quality focus on examining the changes in the amount of data over a specific amount of time from the viewpoint of algorithm designers, big data developers, and data scientists.
Environment and Constraints	The big data pipeline. The factors and parameters that should be understood are resource factors, process factors, methods, and tools.

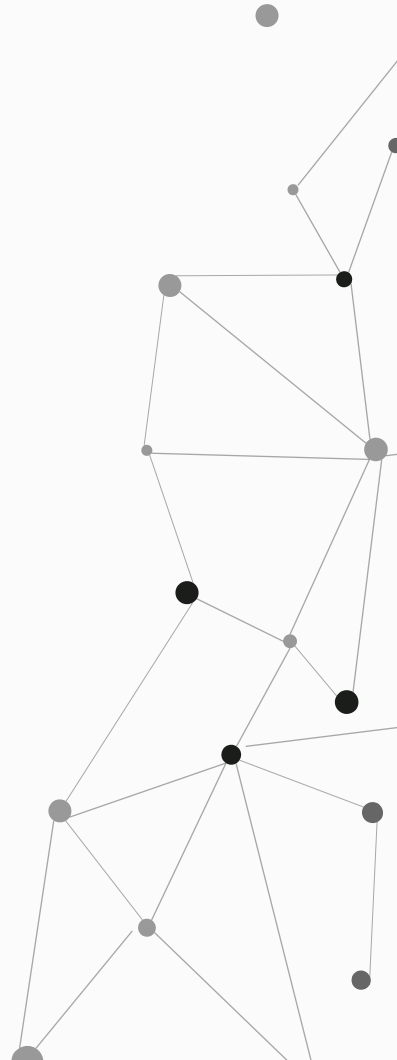


05

Operationalized Goal: Label and Description	**OG #3**: Heterogeneity of the big dataset [3]. Determine the degree of data organization. What's more, it refers to differences in terms of data types, emphasizes different data sources, and emphasis on the diverse function of data elements [4].
Corresponding Measurement Goal Label	**MG #3**: Big Data Variety.
Object of Interest	The big data information product and resource.
Purpose	Big data variety is a measure of the richness and fruitfulness of the data representation. The purpose of this process is to analyze and handle data from various sources in different forms of structures. Big data variety ought to empower the big data system, add complexity, and it is one of the obstacles for effectively using huge dataset and making decision.
Quality Focus, Perspective	The quality focus on examining the changes and the diversity of the structure and the format of the data. The perspective is from the point view of the manager, engineer, and process improvement team.
Environment and Constraints	The big data program and the machine learning pipeline. The factors and parameters that should be understood are application factors, resource factors, methods, tools, and constraints.

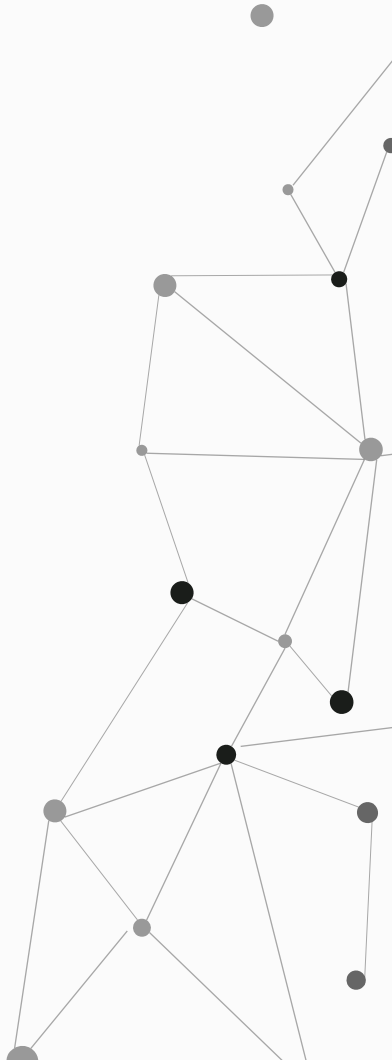


Operationalized Goal: Label and Description	**OG #4**: Focus on and improve data quality and accuracy and defines how data can be trusted when the important decision needs to be made regarding the collected data in the implementation.
Corresponding Measurement Goal Label	**MG #4**: Big Data Veracity.
Object of Interest	The big data information resource and task.
Purpose	Big data veracity is to evaluate whether the big data is categorized as good, bad, or undefined, which is caused by data inconsistency, incompleteness, ambiguity, latency, deception, and approximations [5]. The purpose of the process is to comprehend the data sensitivity in order to protect data and comply with the regulatory requirements [6].
Quality Focus, Perspective	The quality focus on examining the level of value and certainty of the big dataset. The perspective of the operationalized goal is from the point of view of the developer, manager, and engineer.
Environment and Constraints	The big data program and the machine learning pipeline. The factors and parameters that should be understood are application factors, people factors, resource factors, process factors, and constraints.



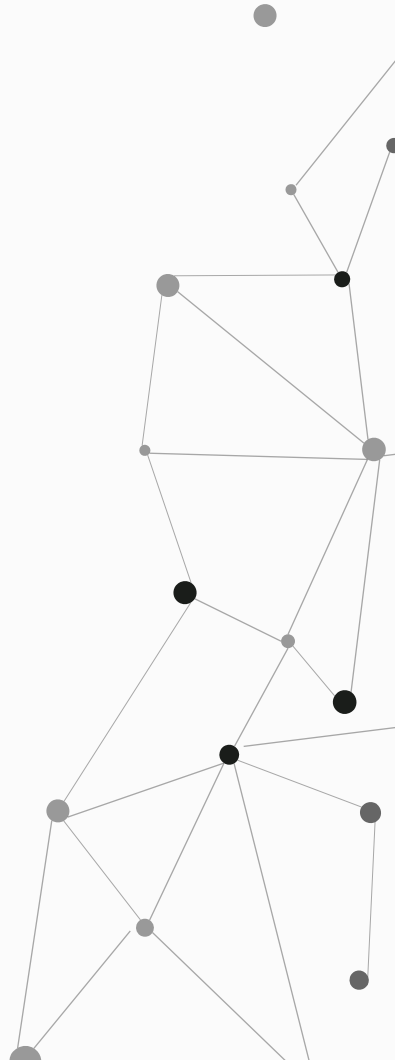
05

Operationalized Goal: Label and Description	**OG #5**: Improve data accuracy and correctness to ensure that data is valid when the status of data changes from exploratory to actionable.
Corresponding Measurement Goal Label	**MG #5**: Big Data Validity.
Object of Interest	The big data information process and activity.
Purpose	Big data veracity is to characterize whether the property of the big data is valid after it is combined from various diverse sources in order to find the presence of hidden relationships among elements within huge Big Data generation sources [7].
Quality Focus, Perspective	The quality focus on examining the degree and quality of correctness of the big dataset in the big data pipeline. The perspective is from the point of view of the developer, manager, and engineer.
Environment and Constraints	The big data program and the machine learning pipeline. The environmental factors and related parameters that should be understood are application factors, resource factors, process factors, methods and tools.



05

Operationalized Goal: Label and Description	**OG #6**: Improve data connectivity and linkage to extract worthwhile data from several diversely connected datasets in the implementation.
Corresponding Measurement Goal Label	**MG #6**: Big Data Vincularity.
Object of Interest	The big data information process, product, resource, task and activity.
Purpose	Big data veracity is to characterize the feasible methods for the aspect of the big data to find the common or semantically equivalent or related attributes to link data sources. Without rigorous semantics to establish linkages, the value of datasets after data integration could be lost [2].
Quality Focus, Perspective	The quality focus on examining the stability and continuity of the big dataset from the perspective of the manager, customer, and senior management.
Environment and Constraints	The big data program and the machine learning pipeline. The environmental factors and related parameters should be understood are application factors, people factors, process factors, customer factors and constraints.



Reference

1. Caballero I, Serrano M, Piattini M. A data quality in use model for big data. In International Conference on Conceptual Modeling 2014 Oct 27 (pp. 65-74). Springer, Cham.
2. Bhardwaj D, Ormandjieva O. Rigorous Measurement Model for Validity of Big Data: MEGA Approach. In25th International Database Engineering & Applications Symposium 2021 Jul 14 (pp. 285-291).
3. Alsaig A, Alagar V, Ormandjieva O. A critical analysis of the V-model of big data. In2018 17th IEEE International Conference On Trust, Security And Privacy In Computing And Communications/12th IEEE International Conference On Big Data Science And Engineering (TrustCom/BigDataSE) 2018 Aug 1 (pp. 1809-1813). IEEE.
4. Moraga C, Moraga MÁ, Calero C, Caro A. SQuaRE-aligned data quality model for web portals. In2009 Ninth International Conference on Quality Software 2009 Aug 24 (pp. 117-122). IEEE.
5. Felderer M, Russo B, Auer F. On testing data-intensive software systems. InSecurity and Quality in Cyber-Physical Systems Engineering 2019 (pp. 129-148). Springer, Cham.
6. Guerra-García C, Caballero I, Piattini M. Capturing data quality requirements for web applications by means of DQ_WebRE. InProceedings of the 2nd International Workshop on Business intelligencE and the WEB 2011 Mar 25 (pp. 28-35).
7. Agrawal, D., Bernstein, P., Bertino, E., Davidson, S., Dayal, U., Franklin, M., & Widom, J. (2011). Challenges and opportunities with Big Data 2011-1.