

SOEN 6611 - SOFTWARE MEASUREMENT: THEORY AND PRACTICE
Project Report on Task 3
WINTER 2022
Course Instructor: Dr. Olga Ormandjieva

Source: SEI Implementing Goal-Driven Measurement course material (adapted).

TEAM 5	
Student #	Name
40163582	Mohammod Suhel Firdus
40157109	Vivekananda Reddy Gottam
40184906	Saswati Chowdhury
40156971	Milesh Kotadia

SOEN6611/W22 Project Step 3 (10 points, due on March 26th): 3V's Success Criteria and Indicators, derived measures and base measures

Step3-Part 1 (6 points): derive 3V's Success Criteria and Indicators

The objective of Part 1 is to develop success criteria and success indicators.

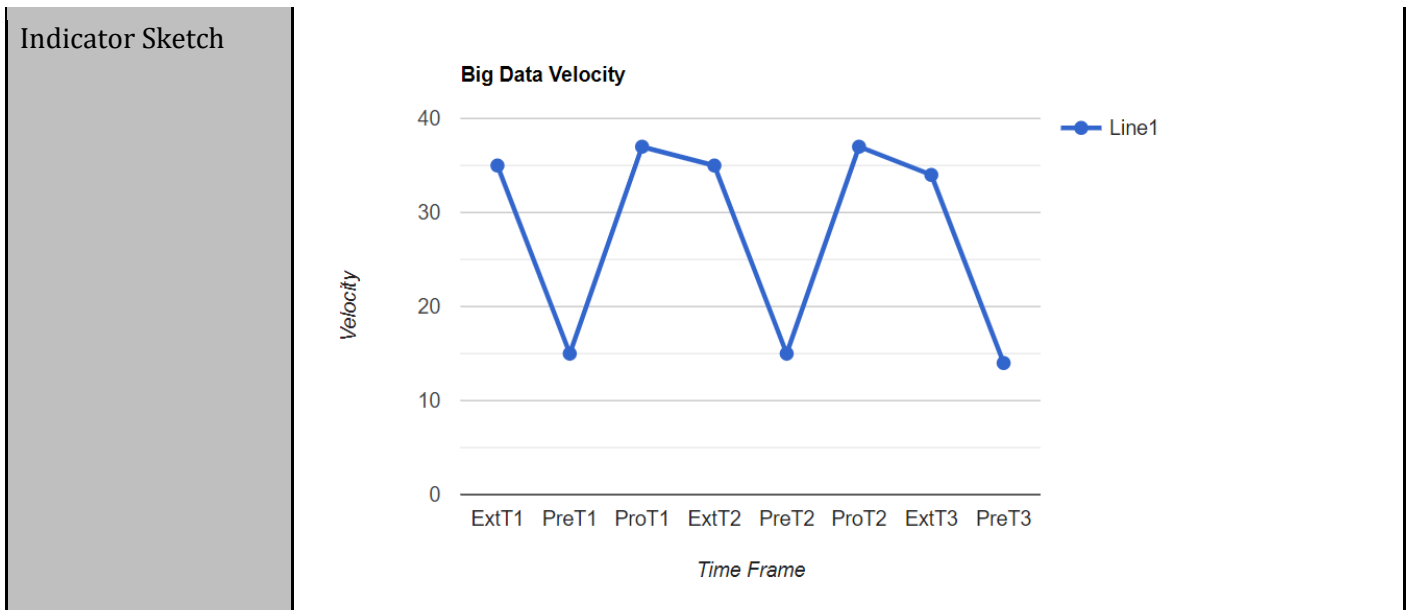
Success (answering the measurement question within the desired timeframe) can only be achieved when certain conditions are in place. indicators that will allow you to answer the questions quantitatively and then communicate the results to others.

Measurement Question Label / Operationalized Goal Label	MG1 - Volume Select the dataset with high-quality Volume.
Success Criteria Label and description	The success criteria of volume: The percentage of data which can be preprocessed increases with the increase of volume of dataset and decreases with the decrease in the volume.
Indicator Label and description	<I1> Mvol: The information content of multiple datasets is specified by the number of information bits in all the records.
Indicator Analysis Model and Interpretation	Indicator Analysis: The volume is divided for each time frame because it is the only source of data that is available. In each time frame, the volume passes through three phases: Extraction, Preprocessing, and processing. Interpretation: We compare the difference between each phase of volume for each time frame.

Indicator Sketch

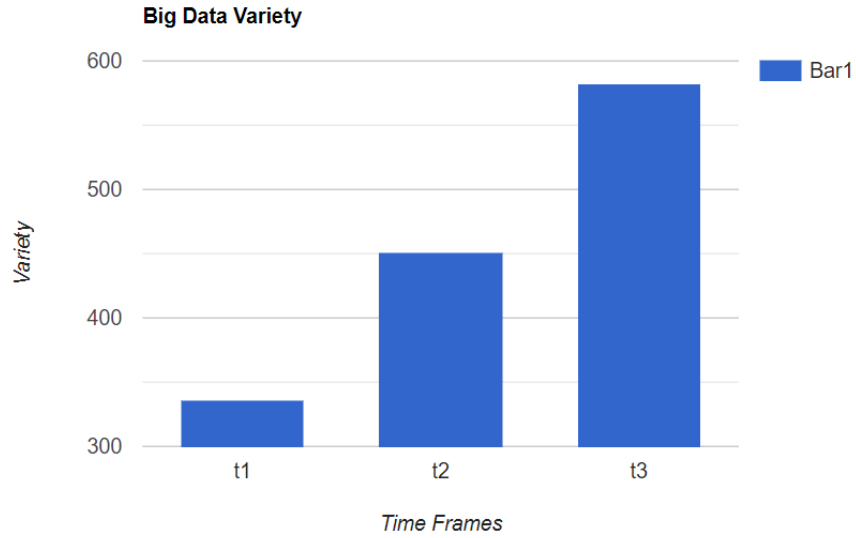


Measurement Question Label / Operationalized Goal Label	MG2 - Velocity Improve the decision by frequently gathering data and processing it faster.
Success Criteria Label and description	The success criteria of velocity: An extreme difference in the velocity of the dataset during different time frames indicates an outdated dataset. The amount of meaningful data which can be processed increases with respect to the increase in volume with respect to time and vice versa.
Indicator Label and description	<I2> Mvel: The rate at which big data volume increases over time (T).
Indicator Analysis Model and Interpretation	Indicator Analysis: In each time frame, the amount of information gain is analyzed after passing the dataset from three phases: Extraction, Preprocessing and Processing. Interpretation: Compare the percentage of changes in the number of insights in order to gain over the range of time period.



Measurement Question Label / Operationalized Goal Label	MG6 - Variety The purpose of Variety is to classify and separate data through categorization and segmentation.
Success Criteria Label and description	When amount of data, the number of records, and the number of datasets increases from previously used dataset
Indicator Label and description	<I3> Mvar: A Mvar (MDS) is expressed as a set of three values (Ndde, Lbd, Nds) aggregated into a single value to indicate the diversity of unique data elements, records, and datasets in a given MDS.
Indicator Analysis Model and Interpretation	Indicator Analysis: The dataset is divided with respect to different time frames to analyze the variety in each time. Interpretation: We will compare the variety in each dataset in different time periods to evaluate the trend for structured data & unstructured data.

Indicator Sketch



Step3-Part 2 (4 points): The objective of Part 2 is to define all measures required to derive your 3V's indicators (Volume, Variety, Velocity) and decide on the achievement of the corresponding operationalized goals.

3.2.1 Identification of the 3V's measures, tracing them to the indicators, their availability and source

Indicator level	Indicators	Formula
I1	Mvol	$Mvol(MDS) = Ndde(MDS) * \log_2(Ndde(NDS))$
I2	Mvel	$Mvel(MDS) = ((Mvol(MDS_{T2}) - Mvol(MDS_{T1})) / Mvol(MDS_{T1}) * 100$
I3	Mvar	$Mvar(MDS) = Ndde(DE) * W_{Ndde} + Lbd(MDS) * W_{Lbd} + Nds(MDS) + W_{Nds}$

Measures					Indicator(s) label		
#	Identification (name of the measure)	Type	Availability	Source*	<I1>	<I2>	<I3>
1	Ndde - Number of Distinct Data Elements	Base	C	Dataset	X		X
2	Lbd: Length of Big Data (Number of Records	Base	A	Dataset	X		X
3	Nds: Number of Datasets	Base	A	Dataset			X
4	Time	Base	A	Dataset		X	
5	Mvol	Derived	B	Dataset	X		
6	Mvel	Derived	B	Dataset		X	
7	Mvar	Derived	B	Dataset			X

Legends:

Type: "Derived" or "Base".

Availability: "A": Already available and collected , "B": Can be derived from other data fairly directly; , "C": Possibly obtained with minor effort; "D": Not available at the moment; "E": Very difficult, if not impossible to obtain at the moment.

Source: Place or tool where data is collected. In the case of base measures, this is obvious; in the case of derived measures, it depends on where the base data is stored after collection.

Indicator (s): Mark an "X" when this measurement is required for each of your indicators.

* <https://www.kaggle.com/anoopkumarraut/most-popular-superhero-tv-shows>

3.2.2 3V's Derived measures: definitions and operationalization

Derived measure or indicator: Volume				
#	Derived measure or indicator Mvol	Formula $\text{Mvol (MDS)} = \text{Ndde (MDS)} * \log_2 ((\text{Ndde (NDS)}))$ Where, Ndde(MD) = Number of Distinct data Element across MDS		
Link with the measurement goal (which goal) MG1 - Increasing the Volume of big data sets		Responsible (who analyzes) Data Scientist	Stakeholder (who uses) Product Owner, Developer	Frequency (when) Before starting the development process of the ML Algorithm (DAY 0). Before each time the dataset changes. (DAY0 + N) WHERE N =no of days when dataset changes
Data source (where the measurement data will be extracted from) Most Popular Superhero TV Shows https://www.kaggle.com/anoopkumarraut/most-popular-superhero-tv-shows		Storage of the result (where data will be stored after the extraction) Local Storage or Any distributed File System	Data interpretation rules The volume in each phase (Extraction, Preprocessing and Processing) can either increase or decrease, indicating a positive or negative association, except we expect the volume to increase to add new data and decrease in order to remove it. Our decision is directly related to the actions we take in the data pipeline.	

Analysis procedure

As in the above formula, data is collected at a particular time and location (for example, a volume measurement may occur in parallel with the data extraction phase and the data preprocessing phase at the same time). During pipeline progression or between time frames at the same stage, the volume of data can be compared.

W.r.t the dataset in analysis:

$$Mvol(MDS) = Ndde(MDS) * \log_2(Ndde(MDS))$$

For T1

mVolT1Ext = 12216

mVolT1Pre = 10537

mVolT1Pro = 7770

For T2

mVolT2Ext = 11994

mVolT2Pre = 10377

mVolT2Pro = 7637

For T3

mVolT2Ext = 11576

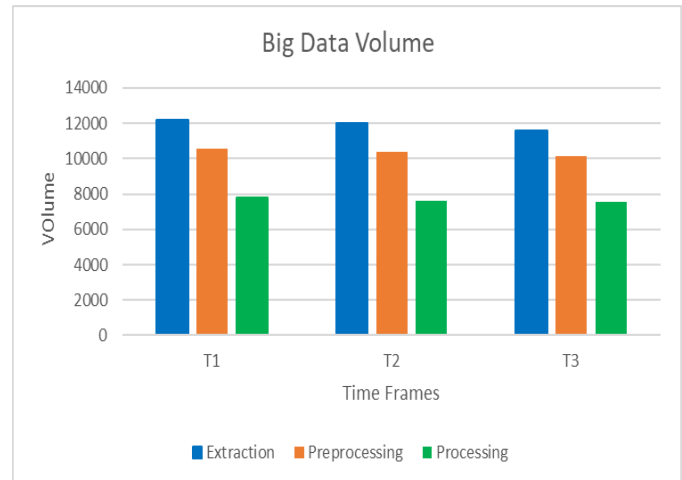
mVolT2Pre = 10125

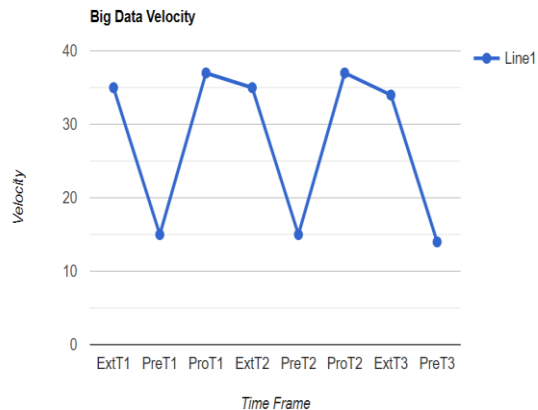
mVolT2Pro = 7526

Potential decision making depending on the results

- 1. Information contents can be compared for multiple datasets using this measurement to determine the actual data that can be processed.**
- 2. These measures help in decision-making for selecting the datasets for further processing and give confidence in the volume of the dataset.**

Presentation of the results (sketch illustrating what it looks like):



Derived measure or indicator: Velocity				
#	Derived measure or indicator Mvel	Formula $\text{Mvel}(\text{MDS}) = ((\text{Mvol}(\text{MDS}_{T2}) - \text{Mvol}(\text{MDS}_{T1})) / \text{Mvol}(\text{MDS}_{T1})) * 100$ <p>Where, MDST1 and MDST2 are the multiple datasets at time T1 and T2 respectively (where T2>T1). Thus, Mvol (MDS) is defined in terms of volume growth over an interval of time (T2-T1) along with the appropriate unit of measure (seconds, minutes, hours, weeks, etc.).</p>		
Link with the measurement goal (which goal) MG3 - Accelerate the Big Data set Velocity		Responsible (who analyzes) Data Scientist	Stakeholder (who uses) Developer, Data scientist	Frequency (when) In each time frame for all three phases - Extracting, preprocessing, Processing
Data source (where the measurement data will be extracted from) Most Popular Superhero TV Shows https://www.kaggle.com/anoopkumarraut/most-popular-superhero-tv-shows		Storage of the result (where data will be stored after the extraction) Local Storage or any distributed File System	Data interpretation rules Mvel value represents a time-dependent change in volume. <ol style="list-style-type: none"> 1. Positive plane in the graph indicates getting more useful information. 2. Negative plane in the graph indicates losing useful information. 3. Straight plane in the graph indicates no gaining and losing information. 	
Analysis procedure In order to compare the volume of big data over time, we need to draw a line graph illustrating the rate of change in Mvel for three consecutive time periods. An incline or decline difference between the phases of the time frames indicates the change rate is not similar. Values at various time frames and The set below are defined as per the following classification {T1Ext,T1Pre,T1Pro,T2Ext,T2Pre,T2Pro,T3Ext,T3Pre,T3Pro } Ndde = {1195,1050,805,1176,1036,793,1140,1014,783} Mvol = {12216,10537,7770,11994,10377,7637,11576,10125,7526}			Presentation of the results (sketch illustrating what it looks like): 	

$Mvel(MDS) = \left(\frac{Mvol(MDS_{T2}) - Mvol(MDS_{T1})}{Mvol(MDS_{T1})} \right) * 100$ <p>The values captured are as follows</p> <p>{35,15,37,35,15,37,34,14}</p>	
<p>Potential decision making depending on the results</p> <p>From the graph, we can determine the rate at which information can be gained or lost. By monitoring regularly, it keeps the data away from becoming outdated by maintaining a similar velocity during a given period and it finds the amount of outdated data if the velocity is not similar in each time frame.</p>	

Derived measure or indicator: Variety				
#	Derived measure or indicator	Formula		
	Mvar	$\text{Mvar (MDS)} = \text{Ndde (DE)} * W_{\text{Ndde}} + \text{Lbd (MDS)} * W_{\text{Lbd}} + \text{Nds(MDS)} + W_{\text{Nds}}$ WNdde : Weight of Ndde (Set to 1/3 by default) WLbd : Weight of Lbd (Set to 1/3 by default) WNds : Weight of Nds (Set to 1/3 by default) Sum of all weights is equal to 1		
Link with the measurement goal (which goal) MG2 - Enhancing Variety in Big Data		Responsible (who analyzes) Data Scientist	Stakeholder (who uses) Developer, Tester	Frequency (when) In each time frame, variety of the dataset is checked in order to get Ndde, Lbd and Nds.
Data source (where the measurement data will be extracted from) Most Popular Superhero TV Shows https://www.kaggle.com/anoopkumarraut/most-popular-superhero-tv-shows		Storage of the result (where data will be stored after the extraction) Local Storage or Any distributed File System	Data interpretation rules Variety comes up in the amount of Ndde, Lbd, and Nds show in our dataset. The change in the values of variety during different time frames is not useful in determining whether the variety in dataset is good or not. This permits us to see the amount of information, records, and datasets present at a look.	

Analysis procedure

We apply the formula present above with weights that are determined by the data practitioner.

By default, all weights are set to 1/4. If for example, the data practitioner would like to make accuracy more important the weight could be increased allowing them to better see changes in that specific measure better

NddeT1= 2694

lbdT1 = 750

NddeT2 = 2652

lbdT2 = 729

NddeT3 = 2056

lbdT3 = 549

Nds = 3

$mvarT1 = ((nddeT1)/3) + (lbdT1/3) + (Nds/2)$

$mvarT2 = ((nddeT2)/3) + (lbdT2/3) + (Nds/2)$

$mvarT3 = ((nddeT3)/3) + (lbdT3/3) + (Nds/2)$

mvarT1 = 336

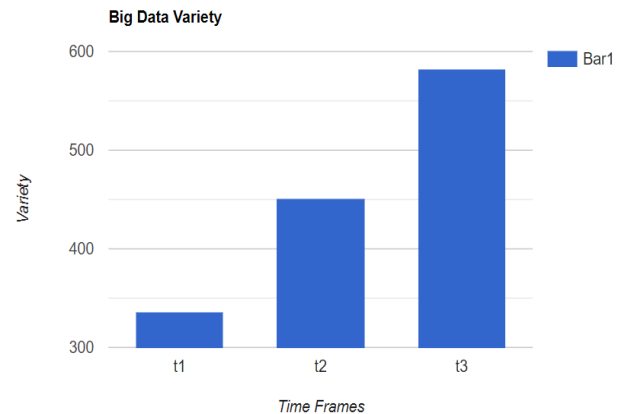
mvarT2 = 451

mvarT3= 582

Potential decision making depending on the results

By analyzing the graph, it can be determined how the amount of variety in the dataset changes during different time periods. By continuously monitoring, it analyzes the changes on the amount of structured, unstructured and semi structured data in every time intervals.

Presentation of the results (sketch illustrating what it looks like):



3.2.3 3V's Base measures: definitions and operationalization

Base measure: Time					
#	Measure (what: entity, attribute) Entity: Dataset Attribute: Time(T)		Scale type Ratio scale	Applicability The given dataset volume is divided into 3 timeframes based on this value	
Who measures?		Source of measurement	Where to store the result	Tool	Time (when to measure)
Developer/Data Scientist		Most Popular Superhero TV Shows https://www.kaggle.com/anoopkumarraut/most-popular-superhero-tv-shows	Local Storage or any distributed File System	Google collab notebook	1) At the time of extracting-processing, and processing the dataset. 2) Repeat every time the dataset changes.
Collection procedure (how to collect the data) The amount of time taken to collect the data by downloading it from Kaggle website (https://www.kaggle.com/anoopkumarraut/most-popular-superhero-tv-shows).			Notes or comments: This measure is generally used to calculate the velocity of data.		

Base measure: Ndde					
#	Ndde : number of distinct data elements Measure (what: entity, attribute) Entity: Dataset Attribute: no of unique elements		Scale type Absolute scale	Applicability Calculates the count of unique elements present in the entire dataset	
Who measures?	Source of measurement	Where to store the result	Tool	Time (when to measure)	
Developer/ Data Scientist			Google Colab	Distinct elements are calculated after dividing the	

	Most Popular Superhero TV Shows https://www.kaggle.com/anoopkumarraut/most-popular-superhero-tv-shows	Local Storage or any distributed File System		dataset volume according to the different time intervals
Collection procedure (how to collect the data) By using the Python code in Google colab online platform to find the distinct elements in the dataset.		Notes or comments: This measure is used to perform the calculation for variety and volume.		

Base measure: Lbd				
#	Lbd : Length of Big Data Measure (what: entity, attribute) Entity: Dataset Attribute: Size	Scale type Absolute Scale	Applicability It gives the actual length of the dataset and can be used to evaluate variety in dataset.	
Who measures? Developer/ Data Scientist	Source of measurement Most Popular Superhero TV Shows https://www.kaggle.com/anoopkumarraut/most-popular-superhero-tv-shows	Where to store the result Local Storage or any distributed File System	Tool Google Colab	Time (when to measure) During each time frame, the length of new dataset is calculated.
Collection procedure (how to collect the data) By using the Python code in Google colab online platform to find the distinct elements in the dataset.		Notes or comments: This measure is to calculate the variety.		

Base measure: Nds				
#	Nds : No of Dataset in Big Data Measure (what: entity, attribute) Entity: Data set Attribute: number of datasets	Scale type Absolute scale	Applicability It counts the number of datasets which are present to analyze the variations coming in each measure during different time period	
Who measures?	Source of measurement	Where to store the result	Tool	Time (when to measure)
Developer/ Data Scientist	Most Popular Superhero TV Shows https://www.kaggle.com/anoopkumarraut/most-popular-superhero-tv-shows	Local Storage or any distributed File System	Google Colab	During each time frame, the count of new dataset is calculated.
Collection procedure (how to collect the data) By using the Python code in Google colab online platform to divide the entire dataset and allocate it to different time frames.		Notes or comments: This measure is to calculate the variety.		

Bibliography:

- [1] Ormandjieva, Olga et al. "Measuring the 3V's of Big Data: A Rigorous Approach." IWSM-Mensura (2020).
- [2] Lecture 11 Notes for performing Step 3 of Project.
- [3] Dave Bharadvaj, "Measurement Framework for Assessing Quality of Big Data (Mega) in Big Data Pipeline".