SOEN 6611

SOFTWARE MEASUREMENT

INSTRUCTOR: DR. OLGA ORMANDJIEVA

PROJECT

TASK 3

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

SUBMITTED ON : 21-October-2022

Declaration:

We, the members of the team, have read and understood the Fairness Protocol and the Communal Work Protocol, and agree to abide by the policies therein, without any exception, under any circumstances, whatsoever.

TEAM-7

SIDDHARTHA NANDA	40200496
BARIQ ISHTIAQ MOHAMMED	40208194
RAJAT KUMAR	40201807
VIKYATH SRINIVISAULU	40218245

TEAM LEADER

SIDDHARTHA NANDA	40200496
Email ID:	siddhartha.nanda@mail.concordia.ca

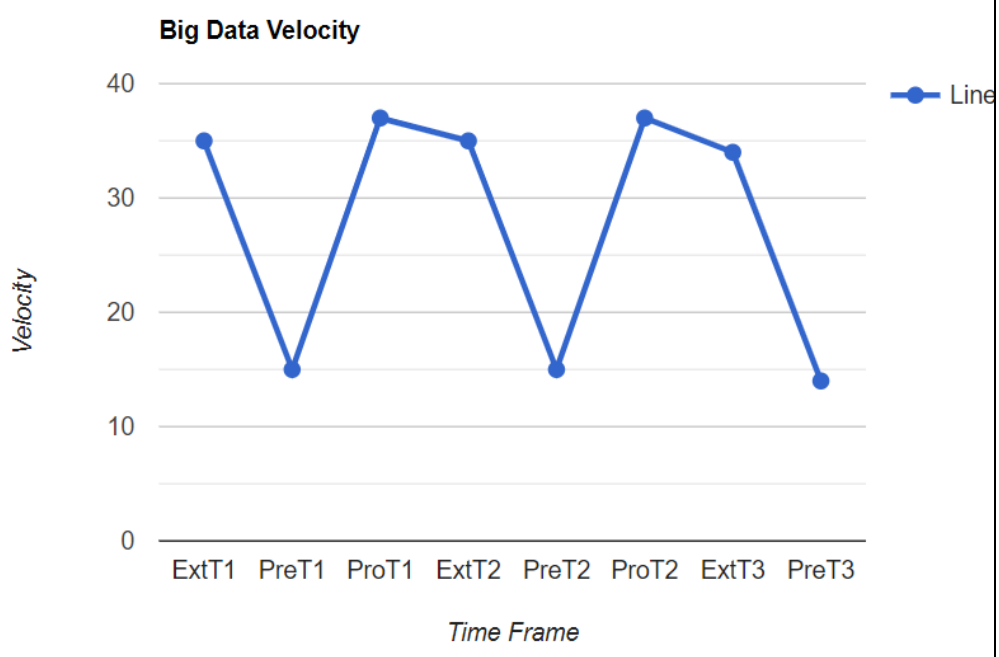
3. Step 3:

3.1. Success Criteria and Indicators for 3 V's:

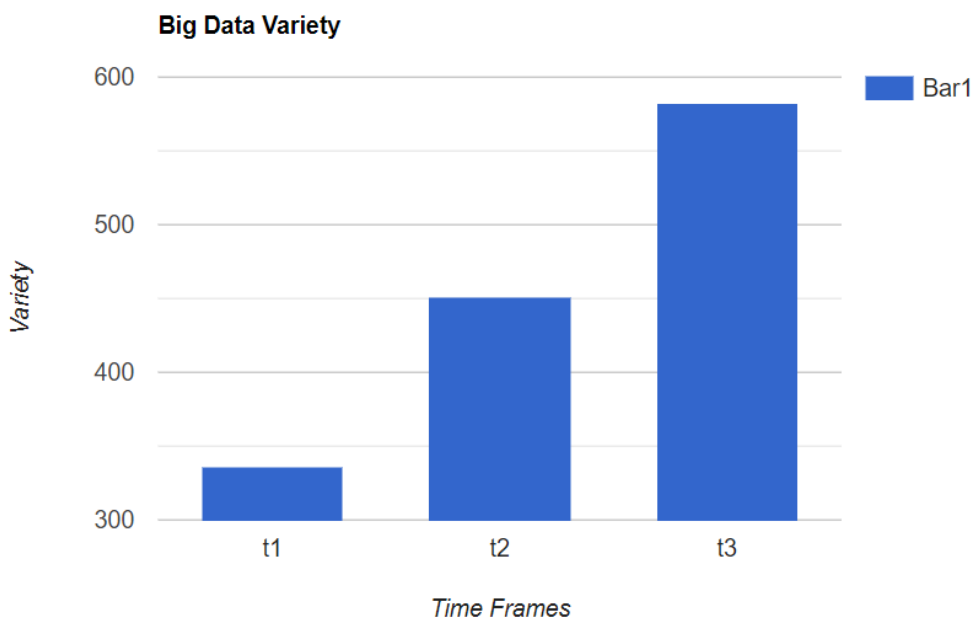
3.1.1. Volume:

Measurement Question Label / Operationalized Goal Label	MG1 - Volume Select the dataset with high-quality Volume.
Success Criteria Label and description	The success criteria for volume are: The percentage of data that can be preprocessed increases with the volume of the dataset and decreases with the decrease in volume.
Indicator Label and description	<I1> Mvol: The quantity of information bits in all the records determines the information content of numerous datasets.
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	

3.1.2. Velocity:

Measurement Question Label / Operationalized Goal Label	MG1 - Velocity Improve the decision-making process by obtaining data more frequently and processing it more quickly.
Success Criteria Label and description	The velocity success criteria are as follows: An excessive discrepancy in the velocity of the dataset between time frames shows that the dataset is out of date. The quantity of useful data that can be processed grows in proportion to the rise in volume over time, and vice versa.
Indicator Label and description	<I2> Mvel: The rate at which big data volume grows over time (T)
Indicator Analysis Model and Interpretation	**Indicator Analysis:** After sending the dataset through three processes in each time frame, the quantity of information acquired is analysed: Extraction, preprocessing, and processing Interpretation: Compare the percentage variations in the number of insights received throughout time periods.
Indicator Sketch	

3.1.3. Variety:

Measurement Question Label / Operationalized Goal Label	MG1 - Variety Variety's goal is to categorise and segment data in order to classify and separate it.
Success Criteria Label and description	The success criteria for variety is: When the amount of data, number of records, and number of datasets rise in comparison to previously utilised datasets
Indicator Label and description	<I1> Mvar: A Mvar (MDS) is a collection of three values (Ndde, Lbd, and Nds) aggregated into a single number to represent the variety of unique data objects, records, and datasets in a specific MDS.
Indicator Analysis Model and Interpretation	**Indicator Analysis:** The dataset is separated into time frames to analyse the variation across each period. Interpretation: The trend for structured and unstructured data will be evaluated by comparing the diversity in each dataset throughout different time periods.
Indicator Sketch	

3.2. Measures and Operationalized Goals

3.2.1. Identification of the 3 V's measures:

a. Volume

Indicator level	Indicators	Formula
I1	Mvol	*Mvol (MDS) = Ndde (MDS) * log_2 ((Ndde (NDS)))*

b. Velocity

Indicator level	Indicators	Formula
I2	Mvel	Mvel(MDS) = ((Mvol(MDS_{T2}) - Mvol(MDS_{T1})) / Mvol(MDS_{T1}) * 100

c. Variety

Indicator level	Indicators	Formula
I3	Mvar	Mvar (MDS) = Ndde (DE) * W_{Ndde} + Lbd (MDS) * W_{Lbd} + Nds(MDS) + W_{Nds}

Measures					Indicators Label		
S.No.	**Identification (Name of the measure)**	**Type**	**Availability**	**Source***	**<l1>**	**<l2>**	**<l3>**
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**

Where:

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.

Indicators:

Mark an "X" when this measurement is required for each of your indicators.

3.2.2. Derived measures definitions and operationalization

<table>
<tr><th colspan="4">Derived measure or indicator: Volume</th></tr>
<tr><td>#</td><td>Derived measures or indicators
Mvol</td><td colspan="2">Formula:

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

Where, Ndde(MD) = Number of Distinct data Element across MDS</td></tr>
<tr><td>Connect to the measurement target (which goal)

MG1 - Increasing the Volume of big data sets</td><td>Responsible:

Data Scientist in charge (who analyses)</td><td>Stakeholder:

Product Owner and Developer, they are the one who uses</td><td>Frequency (when) :

a. Before beginning the ML Algorithm development procedure (DAY 0).
b. Before each update to the dataset.
c. (DAY0 + N) WHERE N = the number of days when the dataset changes</td></tr>
<tr><td>Data source (where the measurement data will be extracted from)

Most popular superhero TV show</td><td>Storage of outcome(where data will be stored after extraction)

Any distributed file system or local storage</td><td colspan="2">Data interpretation rules

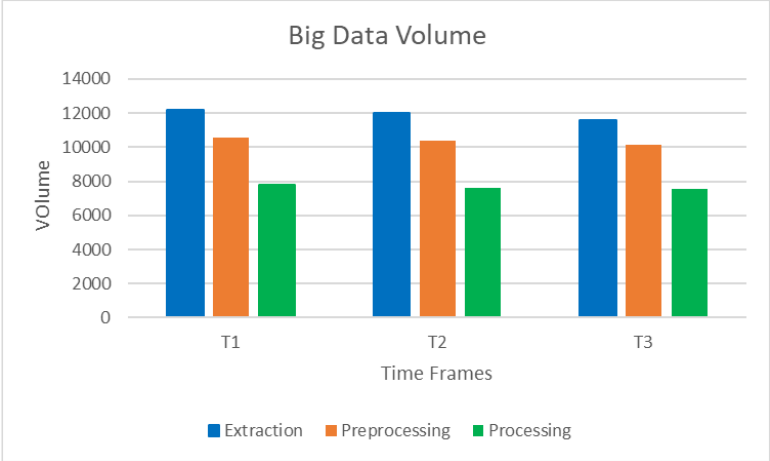
The volume in each step (extraction, preprocessing, and processing) can rise or fall, suggesting a positive or negative connection, with the exception that we anticipate the volume to increase when adding new data and decrease when removing it. Our choice is closely tied to the data pipeline steps we do.</td></tr>
<tr><td colspan="2">Analysis Procedure:

Data is gathered at a certain time and location, as in the preceding formula (for example, a volume measurement may occur in parallel with the data extraction phase and the data preprocessing phase at the same time). The amount of data may be compared throughout pipeline development or between time frames at the same level.

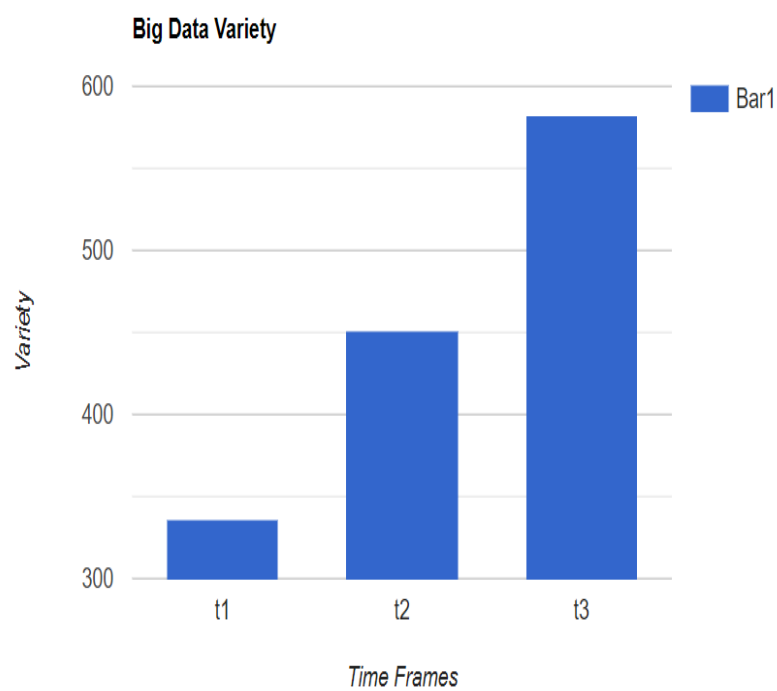
With respect to dataset in analysis:

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

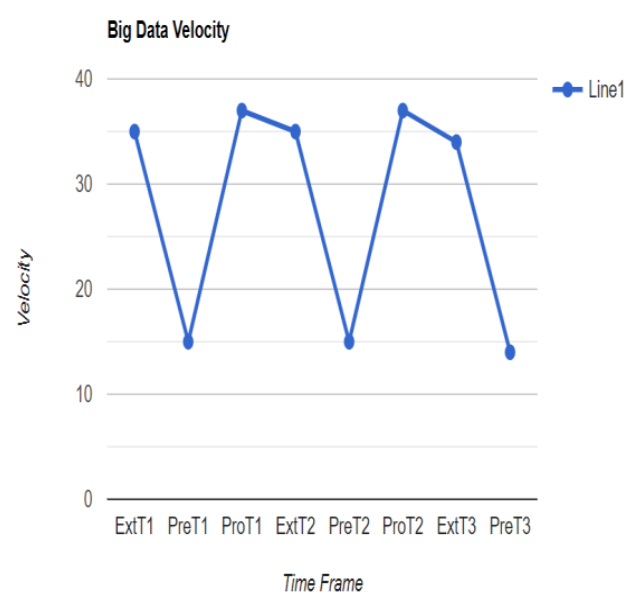
For T1
mVolT1Ext = 12216
mVolT1Pre =10537
mVolT1Pro = 7770

For T2
mVolT2Ext = 11994
mVolT2Pre = 10377</td><td colspan="2">Results presentation (sketch depicting how it looks):
 
</td></tr>
</table>

mVolT2Pro = 7637 **For T3** mVolT2Ext = 11576 mVolT2Pre = 10125 mVolT2Pro = 7526	
Potential decisions based on the outcomes 1. Using this measurement, the information contents of several datasets may be compared to identify the real data that can be processed. 2. These measurements aid in decision-making for picking datasets for further processing and provide confidence in the dataset's volume.	

Derived measure or indicator: Variety			
#	Derived measures or indicators **Mvar**	Formula: **Mvar (MDS) = Ndde (DE) * W_{Ndde} + Lbd (MDS) * W_{Lbd} + Nds(MDS) + W_{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**	
Connect to the measurement goal (which goal) MG2 - Enhancing Variety in Big Data	Responsible: **Data Scientist** in charge (who analyses)	Stakeholder: **Developer and Tester,** they are the one who uses	Frequency (when) : **The variety of the dataset is evaluated in each time frame to provide Ndde, Lbd, and Nds.**
Data source (where the measurement data will be extracted from) **Most popular superhero TV show**	Storage of outcome(where data will be stored after extraction) **Any distributed file system or local storage**	Data interpretation rules **Mvel describes a volume change that occurs over time.** **The quantity of Ndde, Lbd, and Nds in our dataset demonstrates variety. The change in variety values over time is not effective in judging whether or not the variety in a dataset is good. This allows us to see the quantity of information, records, and datasets available at a glance..**	
Analysis Procedure: We use the above-mentioned algorithm with weights specified by the data practitioner. All weights are set to 1/4 by default. If, for example, the data practitioner wants to prioritise accuracy, the weight might be increased, allowing them to observe changes in that specific metric more clearly. NddeT1= 2694 NddeT2 = 2652 NddeT2 = 2056 lbdT1 = 750 lbdT2 = 729 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		Results presentation (sketch depicting how it looks):  	

Potential decision making depending on the results **It is possible to tell how the quantity of diversity in the dataset evolves over time by evaluating the graph. It evaluates changes in the volume of organised and unstructured data by continually monitoring.**	

Derived measure or indicator: Velocity			
#	Derived measures or indicators **Mvel**	Formula: **Mvel(MDS) = ((Mvol(MDS_{T2}) - Mvol(MDS_{T1})) / Mvol(MDS_{T1}) * 100** 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.).	
Connect to the measurement goal (which goal) MG3 - Increase Big Data Set Velocity	Responsible: **Data Scientist** in charge (who analyses)	Stakeholder: **Product Owner and Developer,** they are the one who uses	Frequency (when) : **In each of the three processes - extraction, preprocessing, and processing.**
Data source (where the measurement data will be extracted from) **Most popular superhero TV show**	Storage of outcome(where data will be stored after extraction) **Any distributed file system or local storage**	Data interpretation rules **Mvel describes a volume change that occurs over time.** **1. A positive plane in the graph shows that more meaningful information is being obtained.** **2. A negative plane in the graph implies that useful information is being lost.** **3. A straight plane in the graph shows that no information is gained or lost.**	
Analysis Procedure: To compare the amount of big data over time, we must create a line graph that depicts the rate of change in Mvel across three successive time periods. A variation in inclination or descent between the stages of the time frames indicates that the change rate is not identical. Values throughout various time periods and The following categorization is used to define the set below. {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} $Mvel\ (MDS) = \left(\frac{Mvol(MDS_{T2}) - Mvol(MDS_{T1})}{Mvol(MDS_{T1})}\right) * 100$ The values captured are as follows		Results presentation (sketch depicting how it looks):  	

{35 ,15 ,37, 35, 15 ,37 ,34 ,14}	
Potential decision making depending on the results **We may deduce the pace at which information is obtained or lost from the graph. By monitoring on a frequent basis, it prevents data from becoming obsolete by keeping a comparable velocity throughout a specified time period, and it detects the quantity of outdated data if the velocity is not similar in each time frame.**	

3.2.3. Base measures definitions and operationalization:

Base measure: Time

#	Measure (what: entity, attribute)	Scale type :	Applicability:
	Entity: Dataset **Attribute: Time(T)**	**Ratio scale**	**Based on this value, the provided dataset volume is partitioned into three periods.**

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) During the extraction-processing and processing of the dataset.** **2) Repeat whenever 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 **The number of different elements in the overall dataset is calculated.**

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**	**After splitting the, distinct components are calculated. dataset volume based on different time periods**

Collection procedure (how to collect the data)	Notes or comments:
To locate the different components in the dataset, use the Python code in the Google collab online platform.	**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 provides the real length of the dataset and may be used to assess dataset diversity.**

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 length of new dataset is calculated.**

Collection procedure (how to collect the data)	Notes or comments:
To locate the different components in the dataset, use the Python code in the Google collab online platform.	**This measure is to calculate the variety.**

Base measure: Nds

#	**Nds : No of Dataset in Big Data**	Scale type **Absolute scale**	Applicability

	Measure (what: entity, attribute) **Entity: Data set** **Attribute: number of datasets**			**It counts the amount of datasets that are available to study the fluctuations in each metric across time.**
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**	**The number of fresh datasets is determined at each time period.**

Collection procedure (how to collect the data)

By dividing the full dataset and allocating it to multiple time frames using Python code in the Google collab online platform.

Notes or comments:

This measure is to calculate the variety.