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

**NOTE: the scope of the Steps 3, 4 and 5 is reduced to only 3 of the V’s operationalized in Step 2 namely, Validity, Vincularity and Veracity.**

**SOEN6611/S22 Project Step 3 (10 points, due before 6pm on July 21st)**

**Objective:** develop and document Success Criteria and Indicators, derived measures and base measures

**Step3-Part 1 (6 points): derive Success Criteria and Indicators (for** **Validity, Vincularity and Veracity**)

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

Success (answering the measurement question within the desired timeframe) that 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.

**For each measurement question related to Validity, Vincularity and Veracity**, develop success criteria that will allow you to answer the measurement questions quantitatively.

**Vincularity**

| Mesurément Question Label / Operationalized Goal Label | MG5 / OG5  What is the vincularity of big data? /  To analyze the vincularity of data set over time and measure the percentage change in vincularity for time periods. |
| --- | --- |
| Success Criteria Label and description | S1: vincularity of each dataset should be greater than or equal to 60% for any given time period and percent change for any given two time periods for vincularity should not drop more than 10%. |
| Indicator Label and description | **I1 : Trend Analysis of Vincularity(*M(vin)*) - Trend of Average traceability of multiple datasets over time.**  A line graph is a unique graph which is commonly used in statistics. It represents the change in a quantity with respect to another quantity. Line graph of vincularity over time will provide the trend of vincularity.  **I2 : Mvin - Trend of Vincularity Percentage change of multiple datasets over time.**  **Description:** M(vin) percentage difference is calculated between two time frames. A bar graph containing these values as data points will indicate if the reduction of vincularity quality in the dataset is less than the accepted value or not. |
| Indicator Analysis Model and Interpretation | Tracebility (DS) :  M(vin) :  M(vin) Percentage Diffference:  **Interpretation**  Vincularity will be calculated for each time period T1, T2, T3 and values will be analyzed and percentage difference between vincularity will be analyzed for time periods.  Vincularity(DS)Ti >= 0.6 can be inferred that 60% percent of the data should be traceable for the machine learning algorithm to give relevant results.  **Vincularity = 1.0 - data is completely traceable.**  Drop in Percentage difference of vincularity <= 10% can be interpreted in way that by what percent has the vincularity changed in T2 as compared with T1.  Percentage change - Negative value : Vincularity has decreased.  Percentage change - Positive value : Vincularity has increased.  Percentage change - Negative value (<=10%) : inspect the data and resolve the inaccuracies.  Trend analysis will show us whether the traceability is increasing or decreasing for both processed and raw extract. Increasing vincularity is a good sign and decreasing vincularity means data elements are not tracebale to source and could degrade the performance of machine learning model.  Percentage difference will give us the quantifiable value at which it is increasing or decreasing for given time periods which help us decide whether to accept that dataset with or without any additional modification. |
| Indicator Sketch | | Time Period | Percentage change Raw Data | Percentage change  Processed Data | | --- | --- | --- | | T1 - T2 | 10.9 | 10.9 | | T2 - T3 | 5.66 | 5.66 |   (This can also be plotted as a bar graph) |

**Veracity**

| Mesurément Question Label / Operationalized Goal Label | MG6 / OG6  What is the veracity of data? / To evaluate accuracy & precision of data and trustworthiness of data source over time and analyze the percentage change of subsequent time interval. |
| --- | --- |
| Success Criteria Label and description | 1. Veracity of the dataset should be more than 80%. It is calculated as the combination of accuracy, completeness, currentness and availability.     *Mver >= 0.8*   1. The Mver percentage diffence calculated between two subsequent timeframes should not be less than -5%   *Mver Precentage Difference*: |
| Indicator Label and description | **I3 : Mver - Trend of Average veracity of multiple datasets over time.**  **Description:** M(ver) is calculated for the group of datasets using accuracy, completeness, currentness and availability. The line graph containing these values as a data point will indicate if the veracity is more than the accepted value or not.  **I4 : Mver - Trend of Veracity Percentage difference of multiple datasets over time.**  **Description:** M(ver) percentage difference is calculated between two subsequent time frames. A bar graph containing these values as data points will indicate if the reduction of veracity quality in the dataset is less than the accepted value or not. |
| Indicator Analysis Model and Interpretation | WAcc = WComp = WCurr = WAvail = 0.25  *Mver Percentage Difference*:  **Interpretation**  Veracity will be calculated for T1, T2 and T3 time periods. This will provide us the required data for analysis and the percentage difference occurred during the progression of the project.  To calculate Mver, derived measures such as Accuracy, Currentness, Completeness and Availability should be calculated as per the aforementioned formulas. The summation of products of these values and their corresponding weights will provide Mver.  Mver(MDS)Ti >= 0.8 can be inferred that 80% of the data meets the required veracity standards and can be used by the machine learning algorithm to provide relevant results.  Using the results obtained with above calculator we can understand and track the veracity of our changing dataset. We can then use these values to calculate the percentage difference of Mver between two subsequent timeframes.  If there is an increase in percentage difference of Mver with time, then it can be concluded that data quality is being increased in terms of veracity and it can be trusted. But If the percentage difference is decreased, then the corresponding addition or removal of data has caused the dataset to less dependent and reliable for the algorithm. The acceptable percentage drop is 5%. Any further below will require review of the changes made in the dataset as this can impact the results produced by the machine learning algorithm.  Mver percentage difference >= -5% |
| Indicator Sketch | | Time Period | Percentage change Raw Data | Percentage change  Processed Data | | --- | --- | --- | | T1 - T2 | 0.2 | 0.14 | | T2 - T3 | 0.16 | 0.06 | | T3 - T4 | 0.07 | 0.05 |     It can be seen that in our sample data the percentage difference is comfortambly above than our threshold value i.e., -5% |

**Validity**

| Mesurément Question Label / Operationalized Goal Label | MG4/OG4:  What is the validity of big data? / To evaluate accuracy and correctness of data from different data sources over time. |
| --- | --- |
| Success Criteria Label and description | Every time different data comes to the pipeline, data might not always be compliant and credible. As the increasing or decreasing data should always be accurate and correct. Validity of multiple datasets should be greater than 90% over a time period.  **Validity =1 means data is 100% accurate and correct**  **Description**: At a given instance of point, we calculate the Compliance and Credibility of datasets, and as we progress with project meaning increase or decrease in data used by the Machine Learning algorithms we calculate compliance and credibility. Value of both measures at T2 should be greater than T1. (Provided T2>T1) |
| Indicator Label and description | **I5:** Mval or Validity  **Description**: Mval is calculated for the group of datasets using credibility and compliance. Weighted average is specified for the credibility and compliance. |
| Indicator Analysis Model and Interpretation | **Indicator Analysis**:  Validity is calculated at the Data Extraction step and at the Data Processing step. The validity of big data sources and subsequent analysis must be accurate if we are using the results for decision making. Valid input data followed by correct processing of the data should yield accurate results.  **Mval(MDS)** = Credibility(MDS)\*WCred+Compliance(MDS)\*WCompli  Where,  **Interpretation**: We calculate the Validity of Datasets at the Data Extraction step and Data Extraction step at given Time T1. We also calculate the validity at both steps at given Time T2. As we deal with more data, the data should be accurate and correct so the validity of data at given steps at Time T2 should always be increasing in value. This will make sure our dataset is accurate and correct. Increasing value of validity means the Data is more accurate and correct. Decreasing value of validity means the data may not be correct and accurate to the standard specified. |
| Indicator Sketch |  |

**Step3-Part 2 (4 points)**: The objective of Part 2 is to define all measures required to derive your V’s indicators (for **Validity, Vincularity and Veracity**) and decide on the achievement of the corresponding operationalized goals.

**3.2.1 Identification of the V’s measures** (for **Validity, Vincularity and Veracity**)**, tracing them to the corresponding indicators, their availability and source**

For each of the **V’s** indicators (for **Validity, Vincularity and Veracity**), identify all required measures (derived and base). The table below will be used to complete each of these measures in sections 3.2 and 3.3. It is also recommended that you review and complete this table after all measures have been defined.

This table therefore gives a good summary of all the measurements to be collected and analyzed.

| **Measures** | | | | | **Indicator(s) label** | | |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **#** | **Identification (name of the measure)** | **Type** | **Availability** | **Source** | **<l1>** | **<l2>** | **<I3>** | **<I4>** | **<I5>** |
|  | **Validity** |  |  |  |  |  |  |  |  |
| 1 | Credibility | Derived | C | Program memory |  |  |  |  | X |
| 2 | Nds\_cr: Credible Datasets | Base | A | Dataset |  |  |  |  | X |
| 3 | Nds: Total number of datasets | Base | A | Dataset | X | X |  |  | X |
| 4 | Compliance | Derived | C | Program memory |  |  |  |  | X |
| 5 | Nrec\_comp: Compliant records | Base | B | Dataset |  |  |  |  | X |
| 6 | Mval : Validity | Derived | C | Program Memory |  |  |  |  | X |
|  | **Vincularity** |  |  |  |  |  |  |  |  |
| 1 | Tracebility | Derived | C | Program memory | X | X |  |  |  |
| 2 | Nds(Number of Datasets) | Base | A | Dataset | X | X |  |  |  |
| 3 | Rec\_Trace - total number of records that are traceable | Base | C | Dataset | X |  |  |  |  |
| 4 | Ldst - total number of occurrence of data elements | Base | B | Dataset | X |  |  |  |  |
| 5 | Mvin : Vincularity | Derived | C | Program Memory | X | X |  |  |  |
|  | **Veracity** |  |  |  |  |  |  |  |  |
| 1 | Currentness | Derived | C | Program memory |  |  | X | X |  |
| 2 | Lbd - number of records in dataset | Base | A | Dataset |  |  | X |  |  |
| 3 | Rec\_acc\_age - number of records within acceptable range of age. | Base | B | Dataset |  |  | X |  |  |
| 4 | Availability | Derived | C | Program memory |  |  | X | X |  |
| 5 | N\_succ\_req - number of successful requests | Base | C | Dataset |  |  | X |  |  |
| 6 | N\_req - total number of requests | Base | C | Dataset |  |  | X |  |  |
| 7 | Completeness | Derived | C | Program memory |  |  | X | X |  |
| 8 | rec\_no\_null : record with no null values | Base | B | Dataset |  |  | X |  |  |
| 9 | Accuracy | Derived | C | Program memory |  |  | X | X |  |
| 10 | H\_acc : Entropy | Derived | C | Program Memory |  |  | X |  |  |
| 11 | H\_max : maximum entropy | Derived | C | Program Memory |  |  | X |  |  |
| 12 | P(j) : number of duplicate items in specific dataset. | Base | B | Dataset |  |  | X |  |  |
| 13 | M(ver) : Veracity | Derived | C | Program Memory |  |  | X | X |  |

***Availability:***

***Type****: "Derived" or "Base".*

*"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.*

**3.2.2** **Validity, Vincularity and Veracity: Derived measures definitions and operationalization**

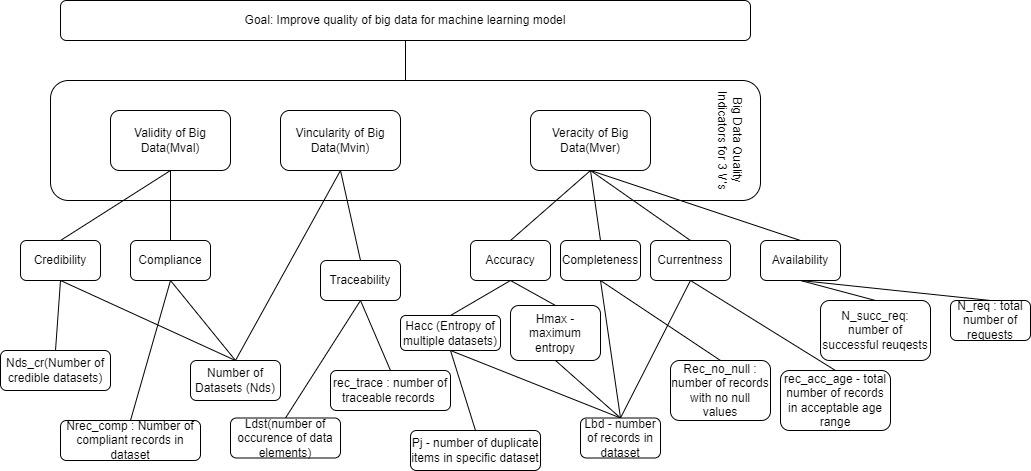
**3.2.3 Validity, Vincularity and Veracity: Base measures definitions and operationalization**

See File : Derived-Base-Measures-Veracity.docx

See File : Derived-Base-Measures-Vincularity.docx

See File : Derived-Base-Measures-Validity.docx

**Measurement Hierachy for 3V’s of Big Data ( Veracity, Vincularity and Validity)**

****