# DAT 325 Project One

### Sup. Smith:

Due to the acquisition of Wayne Enterprises by Bruce, Inc. and the joining of data from the two companies it is imperative that the new data be reviewed for quality. As you are aware, Sebastian-Coleman (2013) stated, “The level of quality of data represents the degree to which data meets the expectations of data consumers, based on their intended uses of the data.” As our internal data consumers are used to receiving data of a certain quality, which we must strive to maintain, this report will serve as a guide to meet this goal.

Data consumers and the associated stakeholders require high quality data so that any information and knowledge derived from the data is trustworthy. To achieve this the six quality dimensions, below, should be adhered to:

* Completeness
* Validity
* Uniqueness
* Timeliness
* Accuracy
* Consistency

Completeness is defined as having all of the necessary parts – for data that includes three ideas: all desired attributes are present, the dataset contains the desired amount of data, and attributes must be populated to the desired density.

Validity is the degree to which the data conforms to business rules and standards.

Uniqueness is the degree to which an entity is repeated within a data set.

Timeliness is the how likely the data is to be delivered according to the prearranged schedule, i.e., data will be available 5 working days from creation.

Accuracy is the measure of the data against a known standard or “to the degree that data correctly represents the “real-life” entities they model” (Mosley, M., & Brackett, M., 2010).

Consistency is the degree to which data conforms to an equivalent set of data.

If these six dimensions are not followed and bad data is allowed to flow through business processes then service quality issues may occur leading to economic ramifications. These come about, for example, by logistics receiving bad sales-velocity information leading to low stocking levels; the organization may face governmental fines for underpaying taxes; loss of customer satisfaction due to shipping product to an incorrect address.

The goals we should adhere to upon joining the data are based on the six dimensions mentioned before:

1. 98% of data entries shall be complete, i.e., no attributes are empty or null, or the offending business process shall be reviewed
2. Duplicate entries shall be removed prior to seeding any models or training neural networks
3. Data consumers shall gain access to their cleaned data no less than 5-business days after creation

These three goals help to achieve the completeness, uniqueness, and timeliness quality dimensions. To accomplish these, the following data quality standards should be adhered to:

1. Data shall be checked for duplicates and any duplicate entries segregated for further analysis
2. Data shall be formatted as below to aid in internal consistency:
   1. Numeric entries are typed as number and not strings
   2. Dates formatted as *MMM-DD-YYYY*
   3. Time is standardized to UTC
   4. Currency is formatted as a string with conversion rate to USD, e.g., “25.16@1.23 USD” for 25.16 British pounds
3. Data that is incomplete, i.e., dimensions are empty and falls below the 98% target, shall be segregated for later review

As the data may contain sensitive customer information, such as, Personal Identifiable Information (PII), usernames and passwords, or tracking data; security procedures must be followed. Moreover, our own company’s sensitive data may be part of the records.

Loss of any customer data could create a loss of credibility in the market leading to lower profitability, a loss of internal data could compromise our ability to do business with creditors and partners. Because of these large stakes, the following precautions shall be followed:

* All data must be categorized as to the level of PII available within the data set using the following levels:
  + Low: no PII present
  + Medium: non-sensitive or indirect PII, e.g., zip code, race, gender
  + High: sensitive PII is present, e.g., credit card information, driver’s license
* All data must be encrypted to at least 128-bit strength (256-bit is preferrable)
* All data labeled *High* shall not be transferred to USB (thumb drives) nor any other removable media
* All data labeled *High* shall be controlled behind a firewall at all times
* Data label *High* must have limited access and only be available to those deemed necessary by the Data Owner
* All data shall be labeled in such a way to comply with GDPR
* Data risk level shall be included in the metadata of any dataset, volume, or store

Because data with a High-risk rating is so sensitive, the principal of least privilege shall be followed. Moreover, data owners must keep a record of who has been granted access to High-risk data. It would also be advisable, but not mandatory, to maintain access logs with the following traits:

* Date and time of access
* IP address of initiating connection
* Username of account accessing records
* Records accessed
* Length of connection

While it is true that encryption, labelling, and control can lengthen analysis time (especially given higher bit-strength encryption) and slow access (reduce timeliness) it is imperative for our continued good reputation in the public and marketplace.

**Citations**

McCallister, E., Grance, T. and Scarfone, K. (2010). *Guide to protecting the confidentiality of personally identifiable information (PII)* (NIST SP No. 800-122). National Institute of Standards and Technology. https://tsapps.nist.gov/publication/get\_pdf.cfm?pub\_id=904990

Mosley, M., & Brackett, M. (Eds.). (2010). *Dama guide to the data management body of knowledge: (Dama-Dmbok Guide)*. Technics Publications, LLC.

Sebastian-Coleman, L. (2013). *Measuring data quality for ongoing improvement: A data quality assessment framework.* Elsevier Science & Technology.