We know that the multi-dimensional view of data mining comprises of four components:

* **Data to be mined**
* **Knowledge to be mined (or: Data mining functions)**
* **Techniques utilized**
* **Applications adapted**

For me the most important element of this view is ***Data to be mined***. As the data would define methods to be utilized to for next steps.

First, we need to identify how we would capture the raw data. Data can be captured from the following sources:

* + Database data (extended-relational, object-oriented, heterogeneous, legacy)
  + data warehouse
  + transactional data
  + stream
  + spatiotemporal (space and time)
  + time-series, sequence
  + text and web
  + multi-media, graphs & social and information networks

Upon retrieval of data, we need to check whether we have all the data points/ features to proceed with our analysis. We also need to identify whether any ETL or cleansing is needed before we start our pre-processing.

If the dataset does not have the necessary features, then we can either derive those features’ values by applying business rules or look for other sources for robust data needed for mining.

Data would dictate what knowledge to be mined. For example, patient diagnosis data would need us to perform characterization analysis for patients’ risk whereas, stock yield data would require predictive analysis.

Then, the structure of the data would require us to follow specific techniques. For some cases, we need machine learning or deep learning to mine the data, for other cases, simple linear regression model would suffice. It all depends on the data we have collected and how it is structured.

Finally, data defines the application for adaptation. For example, we cannot use healthcare data for stock market analysis and vice versa.

Reply 1:

Hello Ben,

Welcome to the class. I would say when it comes to techniques (clustering being one), it may vary depending on the knowledge that you need to mine for your organization and/or industry which is ultimately driven by the data you have on hand. Clustering is indeed one of the most used techniques when it comes to data with demographic, biotechnical and financial dimensions. At the same time, classification and predictive analysis are widely used as well. For example, my organization (healthcare insurance) uses both clustering and predictive analysis for product development and classifications for categorization of sensitive data within our data marts and for fraud detection as well. Then again, when it comes classification we use multiple methods like logistic regression, random forest classifier, KNN, naïve bayes, SVM, XGBoost and deep learning in parallel. The techniques and underlying methods are pretty straight forward and easy to learn in python as the steps are pretty similar. But all boils down to the quality of data that needs to be mined which is our biggest challenge as we get our data from a plethora of sources. So, you need to be very efficient at learning the definitions of the data and preform cleansing if you want to be better at data mining. Best of luck!

Reply 2:

Hello Kusum,

Welcome to the class. Thanks for the interesting read. In an ideal world, we would be performing data mining on data which is landed or loaded into a data warehouse. But unfortunately, that is not the case always. You might have to start the data mining process with data in the raw format like fixed width, txt, psv, csv, json, xml, HL7 etc. or data gathered via web scraping or stream data. You then have to perform ETL to parse the data into a tabular format and perform data cleansing which is a very cumbersome process. I do not want to overwhelm you but to be honest this part is pretty challenging and at the same time very interesting. And it is very satisfying when you get the final clean dataset. What I am looking forward to in this course is learning text analytics because of its growing demand. And Python is a great tool in achieving all these in a pretty simple fashion. Best of luck!