**NBD(NIST Big Data) Requirements WG Use Case Template Aug 11 2013**

|  |  |  |  |
| --- | --- | --- | --- |
| **Use Case Title** | | Statistical Relational AI for Health Care | |
| **Vertical (area)** | | Healthcare | |
| **Author/Company/Email** | | Sriraam Natarajan / Indiana University /natarasr@indiana.edu | |
| **Actors/Stakeholders and their roles and responsibilities** | | Researchers in Informatics, medicine and practitioners in medicine. | |
| **Goals** | | The goal of the project is to analyze large, multi-modal, longitudinal data. Analyzing different data types such as imaging, EHR, genetic and natural language data requires a rich representation. This approach employs the relational probabilistic models that have the capability of handling rich relational data and modeling uncertainty using probability theory. The software learns models from multiple data types and can possibly integrate the information and reason about complex queries. | |
| **Use Case Description** | | Users can provide a set of descriptions – say for instance, MRI images and demographic data about a particular subject. They can then query for the onset of a particular disease (say Alzheimer’s) and the system will then provide a probability distribution over the possible occurrence of this disease. | |
| **Current**  **Solutions** | **Compute(System)** | | A high performance computer (48 GB RAM) is needed to run the code for a few hundred patients. Clusters for large datasets |
| **Storage** | | A 200 GB – 1 TB hard drive typically stores the test data. The relevant data is retrieved to main memory to run the algorithms. Backend data in database or NoSQL stores |
| **Networking** | | Intranet. |
| **Software** | | Mainly Java based, in house tools are used to process the data. |
| **Big Data  Characteristics** | **Data Source (distributed/centralized)** | | All the data about the users reside in a single disk file. Sometimes, resources such as published text need to be pulled from internet. |
| **Volume (size)** | | Variable due to the different amount of data collected. Typically can be in 100s of GBs for a single cohort of a few hundred people. When dealing with millions of patients, this can be in the order of 1 petabyte. |
| **Velocity**  **(e.g. real time)** | | Varied. In some cases, EHRs are constantly being updated. In other controlled studies, the data often comes in batches in regular intervals. |
| **Variety**  **(multiple datasets, mashup)** | | This is the key property in medical data sets. That data is typically in multiple tables and need to be merged in order to perform the analysis. |
| **Variability (rate of change)** | | The arrival of data is unpredictable in many cases as they arrive in real-time. |
| **Big Data Science (collection, curation,**  **analysis,**  **action)** | **Veracity (Robustness Issues, semantics)** | | Challenging due to different modalities of the data, human errors in data collection and validation. |
| **Visualization** | | The visualization of the entire input data is nearly impossible. But typically, partially visualizable. The models built can be visualized under some reasonable assumptions. |
| **Data Quality (syntax)** | |  |
| **Data Types** | | EHRs, imaging, genetic data that are stored in multiple databases. |
| **Data Analytics** | |  |
| **Big Data Specific Challenges (Gaps)** | | Data is in abundance in many cases of medicine. The key issue is that there can possibly be too much data (as images, genetic sequences etc) that can make the analysis complicated. The real challenge lies in aligning the data and merging from multiple sources in a form that can be made useful for a combined analysis. The other issue is that sometimes, large amount of data is available about a single subject but the number of subjects themselves is not very high (i.e., data imbalance). This can result in learning algorithms picking up random correlations between the multiple data types as important features in analysis. Hence, robust learning methods that can faithfully model the data are of paramount importance. Another aspect of data imbalance is the occurrence of positive examples (i.e., cases). The incidence of certain diseases may be rare making the ratio of cases to controls extremely skewed making it possible for the learning algorithms to model noise instead of examples. | |
| **Big Data Specific Challenges in Mobility** | |  | |
| **Security & Privacy**  **Requirements** | | Secure handling and processing of data is of crucial importance in medical domains. | |
| **Highlight issues for generalizing this use case (e.g. for ref. architecture)** | | Models learned from one set of populations cannot be easily generalized across other populations with diverse characteristics. This requires that the learned models can be generalized and refined according to the change in the population characteristics. | |
| **More Information (URLs)** | |  | |
| **Note:** <additional comments> | | | |