**2.**    **Final Report:** A maximum of 7-10 pages (technical) report with three sections:

* Section 1: Introduction (1-1.5 page)
  + Problem Domain: The area or application domain of your project
  + Problem Description (Statement): Concise description of the problem that your team will be solving.
* Section 2: Data Preprocessing (2 pages, maybe less depending on the quality of your data)
  + Describe your data sets, their sources and quality.
  + Discuss in detail how you acquired, wrangled and cleaned your data
  + If you needed to integrate data from different sources, describe the data integration techniques you used to determine relationships between disjoint datasets.
* Section 3: Implementation and Analysis (6-7 pages)
  + Define modeling and/or machine learning techniques you used to solve the problem.
  + Describe the process of training and tuning the machine learning techniques you used.
  + Discuss how you evaluated the performance of your model.
  + If you tried several machine learning techniques, compare the performance of these techniques and discuss why one performed better or worse than the others.

Section 1: Introduction:

Problem Domain:

Our project is in the domain of natural language processing. The broad problem we sought to address was how to judge the semantic intent of language irrespective of its syntactic structure. This is illustrated through a simple example:

Sentence 1: "Is breast cancer preventable?"

Sentence 2: "How do I prevent breast cancer?”

Looking at the two questions we can quickly determine that the two questions do not have the same intent. Sentence 1 asks whether breast cancer is preventable. Likewise, Sentence 2 pre-supposes that breast cancer is preventable, and asks what measures can be taken to prevent breast cancer. The intent of the questions is clearly different. Natural language processing studies in part the mechanisms by which our brains determine that the two questions are different, and how these mechanisms can be reproduced in computers. A naïve approach would be to look at the content words (i.e. nouns, verbs, etc.) and see if they match up between the two sentences. Both contain the phrase “breast cancer” and the prefix “prevent”. Using this measure would produce the wrong classification because it ignores how syntactic structure shapes semantic meaning. Say we remove the content words from the sentences. We are left with:

Sentence 1: "Is *noun* *adjective*?"

Sentence 2: "How do I *verb* *noun*?”

Without knowing the content words, we can already see that the questions are different. This illustrates the point that the semantic intent of a question is not separable from its syntax. The grammatical structure of language shapes the meaning of the language.

Spoken and written language is words conveying meaning chained together by formal rules of grammar. The rules of grammar are a set of axioms we can use to generate meaningful sentences (theorems). Natural language processing is highly effective because we as humans have become very proficient at teaching computers to work within strong axiomatic systems. Using techniques from neural networks to probabilistic graphical models computers computers have become much better at learning the rules of the game. The major obstacle in this challenge, discussed in further detail in Section 2, is that the grammatical rules of language are only used as a loose guideline. People have largely abandoned grammar on the Internet.

Problem Statement:

Our project is the Quora Question Pairs challenge on the Kaggle competition platform. The goal of the challenge is to accurately classify whether a pair of questions “share the same intent”. Rephrased, we are trying to design a classifier that matches human assigned binary classification judging whether two questions have the same intent.

The term “intent” derives from the Latin term for “purpose”.