## Storypoint **Prediction**

## Problem Statement:

In modern agile development settings, software is developed through repeated cycles (iterative) and in smaller parts at a time (incremental), allowing for adaptation to changing requirements at any point during a project’s life. A project has a number of iterations (e.g. sprints in Scrum). Each iteration requires the completion of a number of user stories, which are a common way for agile teams to express user requirements.

There is thus a need to focus on estimating the effort of completing a single user story at a time rather than the entire project. In fact, it has now become a common practice for agile teams to go through each user story and estimate its “size”. Story points are commonly used as a unit of measure for specifying the overall size of a user story.

## Problem Formulation:

**Input:** A string of length that contains a story’s name and description . For each story, a set of text embeddings that contains features extracted from has been provided.

**Output:** A natural number Passociated with the storypoint of that user story.

## Task Description

Your task is to build a machine learning model to predict for given corresponding . The model will be trained to perform predictions on new data.

* **Main Deliverables:** A ML-based technique for estimating storypoint based on that story’s name and description.
* **Stretch Deliverables:** 
  + A detailed analysis on dataset using data exploration techniques such as correlation analysis
  + A comparison between different ML models for Storypoint Prediction
* **Evaluation Metrics:**
  + F1-Score
  + Precision
  + Recall

## **Dataset Information:**

**Text Embeddings.** Text embeddings are a way to convert words or phrases from text into a list of numbers, where each number captures a part of the text's meaning. The dataset has been preprocessed and converted into two kinds of text embeddings. You can choose to work with either of them or both:

* **Doc2Vec:** Input strings are transformed into fixed-length vectors of size 128. These vectors capture the semantic meaning of words and their relationships within a document.
* **Look-upTable:** Input strings are transformed into fixed-length vectors of size 2264. These vectors are obtained via transforming each word in the input strings into an identifier number, then padded to the length of the longest sample.

**Dataset Structure & Format:** Storypoint Estimation Dataset is stored in 3 folders labeled ‘raw data’, ‘look-up’, and ‘doc2vec’. Within each folder are 3 CSV files for training, testing, validation. Each csv file has the following columns:

* issuekey : The unique identifier for a story.
* storypoint: The correct number of storypoint.
* An embedding column (embedding or doc2vec) contains text embedding vectors. The raw data csv will not have this and instead contain two columns with story name and description.

## Acknowledgments:

This dataset is part of a study conducted to facilitate research in effort estimation for agile development. The data has been carefully curated from 6 large open source projects in 9 repositories namely Apache, Appcelerator, DuraSpace, Atlassian, Moodle, Lsstcorp, Mulesoft, Spring, and Talendforge.