**Context -** EKG files are a core source of information on identifying heart related conditions. Most feature extraction mechanisms focus on retrieving core information in an efficient form. Medical practitioners have pre identified health related features (distance between key parts of the waves, eg. QRS complex) and technologists have focused on efficient extraction based on downsampling the data through mathematical representations. We instead wish to reexamine the area by extracting the lowest level features of the EKG and use those as inputs into ML / NN techniques to extract features or make predictions.

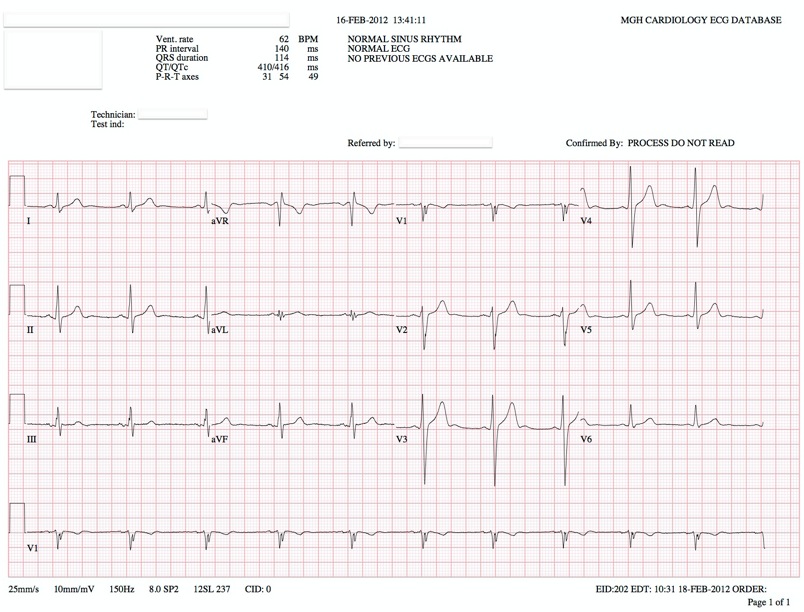
**Goal** - develop a modular software approach to extract low level features from EKG files (images / PDFs). The software should be able to operate on a single file or directories of files in an automated way. The output of the execution should be a csv or JSON file that can be imported easily into other environments. The exact format of the output should be pluggable to support later changes (ie varying input images that may have more or less data in them)

**Assumptions**

* software must be easily deployable without any external dependencies (non cloud based).
* no dependency on commercially licensed software
* use a relevant mainstream programming languages (Java, C/C++, Python, R)
* the program can be pointed at a file(s) or at a director (and will process every file in that directory), output is written to wherever the program is run from or configurable directory.

**Capabilities / Process Description**

* The EKG Document - can be in PDF or image file. (sample file linked here). It consists of two primary parts:
  + The EKG image
    - EKGs have 12 leads (1,2,3, aVR,aVF,aVL,v1,v2,v3,v4,v5,v6) and a rhythm strip. Standard EKG are plotted on a grid of squares 1mm x 1mm in length. Each horizontal grid represents .04 seconds (25 grids = 1 second) and each vertical grid represents amplitude (deflection) and equal .1 mV (10 grids = 1mV).
    - There are four rows of data. The first three rows each contain information from 4 separate leads (ie its not a continuous row, eg first row has a segment of data from leads I, aVR, V1 and V4). The bottom row is the rhythm strip and is continuous (normally uses the data from V1)
    - The far left of each row is a square box that represents the legend of the data (1mV high and .2 seconds wide) normally this will be 10 grid squares high and 5 grid squares wide if it matches the standard format described above. This indicates the units of the graph and that readings are about to begin. It will only be necessary to read / use this data if we start processing EKG that are not formatted in the standard units (this will likely not be needed immediately, but will in the future)
  + Surrounding text
    - surrounding the image is information about the patient or the grid legend. At this time we do not expect to need information from this area (later if the filename does not contain a Medical Record Number, we may need to obtain it from this area of the document.
* Processing / Output
  + the developed software should:
    - identify the area of the PDF or image that the EKG is on (the grid square)
    - assume the standard width / height of the starting box (ie that it is 10 squares in length high and five squares in width)
    - identify the starting x coordinate for each lead (I/II/III will have the same starting x, aVR, aVF, aVL will have the same x, V1/V2/V3 will have the same x, V4,V5,V6 will have the same x)
    - determine the right Y coordinate for each of the four rows of data (value at the center of each of the ~250-300 grids per row)
    - Sample Input



* + - Output format
      * 1 row of data.
      * Columns
        + (1) Rowname (assume this will come from the filename, later it may come from the text in the whitespace of the document)
        + (12) each leads starting x column (one value per lead)
        + (12) ending x for each lead (one value per lead)
        + (4 \* ~265) number of small grids X number of rows

should be named x0n…x3n (where n is the grid #)

* Future considerations
  + The file given here should be assume to be the standard format. In the future we will assume we will get other formats that are very similar, but different. For example an image without surrounding text, an EKG with more or less grid squares (different sizes), or the square box on the right, or different dimensional values for each grid square. We do not need to build immediately for all these possibilities but should expect future enhancements related to them.
  + We are extracting the lowest level data, in the future we may want to extract higher level constructs. We expect this will be through manipulation of the output of these low level features vs rereading the image file.