**Building MIMIC DataBase & Using MIMIC\_Extract for Preprocessing & Cleansing**

0- Download the MIMIC data form Physionet. Extract it or keep it compressed (there are different scripts that can operate on each case). This will take 40+ GB

1- Install: git, Anaconda, and postgressql-10.15 (newer versions of postgressql have problems with large files>4GB).

2- Connect Anaconda to the CMD by adding its path (up to the folder with ‘conda’) to the “Path” in Environment System Variables. (or do that on linux via the terminal PATH=$PATH:/home/user\_name/anaconda3/bin).

3- Create the database to hold the data:  
 3a- download the codes (you need \*.sql’s only) from: https://github.com/MIT-LCP/mimic-code/tree/master/buildmimic/postgres

3b- follow the instructions in:

<https://mimic.physionet.org/tutorials/install-mimic-locally-windows/> (you may go for linux/Mac installation as well)

NOTE: when writing the directories, replace \ with / for windows (it is not automatically corrected depending on the OS, it just throws an error as: access denied).

This step requires about 47GB for the database + 26GB for indices. So allow for about 100GB. It may keep running for 4 hours or something.

4- Follow these instructions for the MIMIC-Extract code to preprocess & filter out data. This may take 5 - 10 hours.

Note: the environment won’t load on Windows (neither can I find the repo’s hash in their log which is said to be working). Therefore, I had to do that on Ubuntu, but this triggered other problems with postgresql which I solved later as explained in this document.

<https://github.com/MLforHealth/MIMIC_Extract>

Problems with postgresql on ubuntu:

* Remove it if it was wrongly installed:   
  dpkg -l | grep postgres (this shows all dependencies)  
  then  
  sudo apt-get --purge remove (then type all the names of the dependencies that appeared from the last command)

Finally, install it (version 10 to avoid file size problems with 13th)  
sudo apt -get install postgresql-10

git clone https://github.com/MIT-LCP/mimic-code.git

and cd to mimic-code/buildmimic/postgres

* Now let’s build the dataset from compressed csv files:

make create-user mimic-gz datadir="$datadir" DBUSER="$USER"

Note, this should solve the authentication failure problem by using the same user as created by postgresql (check the mimic-code repo in the troubleshooting). Don’t forget to paste the data directory (that command is for the compressed dataset).