<https://docs.google.com/drawings/d/1U5vZ0J-902Amc-CYsszRy13-N8KNXtIClTRE8xGo3kQ/edit>

What I did

* Aug 14, 2023:
  + Worked on DLATK tutorial
* Aug 2, 2023:
  + Dataset research: [L Factor Dataset Findings](https://docs.google.com/spreadsheets/d/1qDjK5AZTRsByyd7s9dEW_QL_1RnSJ0hsd9OsoPRePLg/edit#gid=0)
* Jul 18, 2023:
  + Completed MySQL tutorial
  + Worked through DLATK tutorial
* Jul 17, 2023:
  + SQL Workbench issues: not able to connect to server
* Jul 16, 2023:
  + Completed linux tutorial
  + Familiarized myself with the hierarchy of cronus
* Jul 14, 2023:
  + Connected to cronus
* Jul 13, 2023:
  + Joined slack channels
  + Created SSH key and sent it to Adi

Blockers: Anything preventing making progress – should always have notified elsewhere as well

* Jul 17, 2023:
  + SQL Workbench issues: not able to connect to server

Next Steps

* **LFactor Meeting Minutes (Nov 7, 2023):**
  + High level information needed:
    - Average word count
    - Standard deviation of word count
    - Range of word count
  + Finish description table
  + Extract n gram feature tables of all datasets
  + Start with PCA on n grams for all datasets
    - CSV file for results of each PCA
    - Loading matrix -> one column actual n grams, another column factor loadings.
    - Visualize which n grams are loading in a positive dimension and negative dimension
  + Word clouds:
    - Top n grams
    - Bottom n grams
* **LFactor Meeting Minutes (Oct 24, 2023):**
  + Features by Dimensionality Reduction Techniques table
  + Extract other feature sets
  + Moving each feature sets with Dimensionality Reduction Techniques - step 0
  + PCA, Exploratory Factor Analysis, Non-negative matrix factorization - step 1
  + CCA - step 2
  + Four topics for each dimensionality reduction techniques (ngrams, BERT)
    - Could train our own LDA topic model
    - Pretrained topic model from HLAB could be better
  + Metrics to compute:
    - Number of speakers in candor
    - Number of utterances (word counts - length of each utterance)
    - Minimum threshold of 100 words per person
    - Mean of word count and std\_dev of word count should be after we take out instances where the word count is less than 100
    - Group\_threshold 100
* **October 10:** Aggregate text to the respective user (dimensionality reduction)
  + CANDOR Corpus:
    - Sensitive data, not to be shared
    - Study of 1000 subjects assigned randomly to partners multiple times and had them conversate
    - Multiple entries of text per person
    - Feature extraction: age, gender
* ~~Get access to slack~~
  + ~~join latentfactors~~
  + ~~join tech\_qs\_and\_dlatk~~
* Get access to Cronus – email your ssh key to Adi (@adigan) in slack (include Ryan)
  + [~~SSH Keys~~](https://www.digitalocean.com/community/tutorials/how-to-set-up-ssh-keys--2)
    - ~~If you don't have an ssh client for windows,~~ [~~64-bit Putty~~](https://www.chiark.greenend.org.uk/~sgtatham/putty/latest.html) ~~is simplest.~~
  + [Linux bash shell commands](https://www.youtube.com/watch?v=oxuRxtrO2Ag)
    - Copying results files from linux machines:   
      ***\* note: do not copy data from hercules unless you are sure it is allowed and safe to be copied off the machine; ask if unsure \****
      * Windows: <https://winscp.net/eng/index.php>
      * Mac/ Linux: <https://askubuntu.com/questions/312697/copy-files-from-remote-ubuntu-to-local-mac>
    - [Linux file permissions](https://ryanstutorials.net/linuxtutorial/permissions.php)
  + [Basic MySQL](https://www.youtube.com/watch?v=yPu6qV5byu4)
    - copying mysql table from one machine to another:  
      ssh [ssh.wwbp.org](http://ssh.wwbp.org/) 'mysqldump REMOTE\_DB TABLE | gzip -c' | zcat | mysql LOCAL\_DB
    - Importing and exporting csvs: <https://github.com/dlatk/usefulScripts>  
      (see csv2mysql.py and mysqlTocsv.bash)
* do [dlatk tutorials](http://dlatk.wwbp.org/tutorials.html) on cronus
  + dlatk video tutorial
  + transformers in dlatk: <http://dlatk.wwbp.org/tutorials/tut_trns.html>
  + clustering and super topics: <http://dlatk.wwbp.org/tutorials/tut_clustering.html>
* Run first version