Child abuse and neglect is social problem that has reached epidemic proportions. The broad adoption of electronic health records in clinical settings offers a new avenue for addressing this epidemic. SAFElab and colleagues are developing an innovative artificial intelligence system to detect and assess risk for child abuse and neglect within hospital settings to prioritize the prevention and reduction of bias against Black and Latinx communities.

Objectives:

Aim 1: To create a taxonomy of risk factors that will assist in the detection of child abuse and neglect (via clinician and primary caregivers domain experts, including nurses, social workers, parents, and legal guardians).
Aim 2: To develop an innovative, clinical decision support algorithm utilizing machine learning for clinicians that will foster objectivity in detecting child abuse and neglect.

Innovation:

1. To approach child abuse and neglect from an interdisciplinary viewpoint empowered by emerging technologies, innovating methodology, and practice.
2. To tackle the concurrent racially related over-detection and under-detection of child abuse and neglect.
3. To formulate a framework with domain expertise, supporting the pre-development of AI systems on medical data and validating the system upon launch.

Methodology

Study Dataset and Population

This study received institutional review board approval at Columbia University. The study used clinical notes from a single pediatric ED in the Northeast region of the United States. The database included 287757 clinical notes of children and youth admitted to the pediatric ED in 2018. Furthermore, we obtained a manual summary of reportings concluded within the ED from the Child Justice Center (CJC). In 2018, CJC conducted 89 reports for child abuse and neglect, and only **75** reports have unique MRN's.

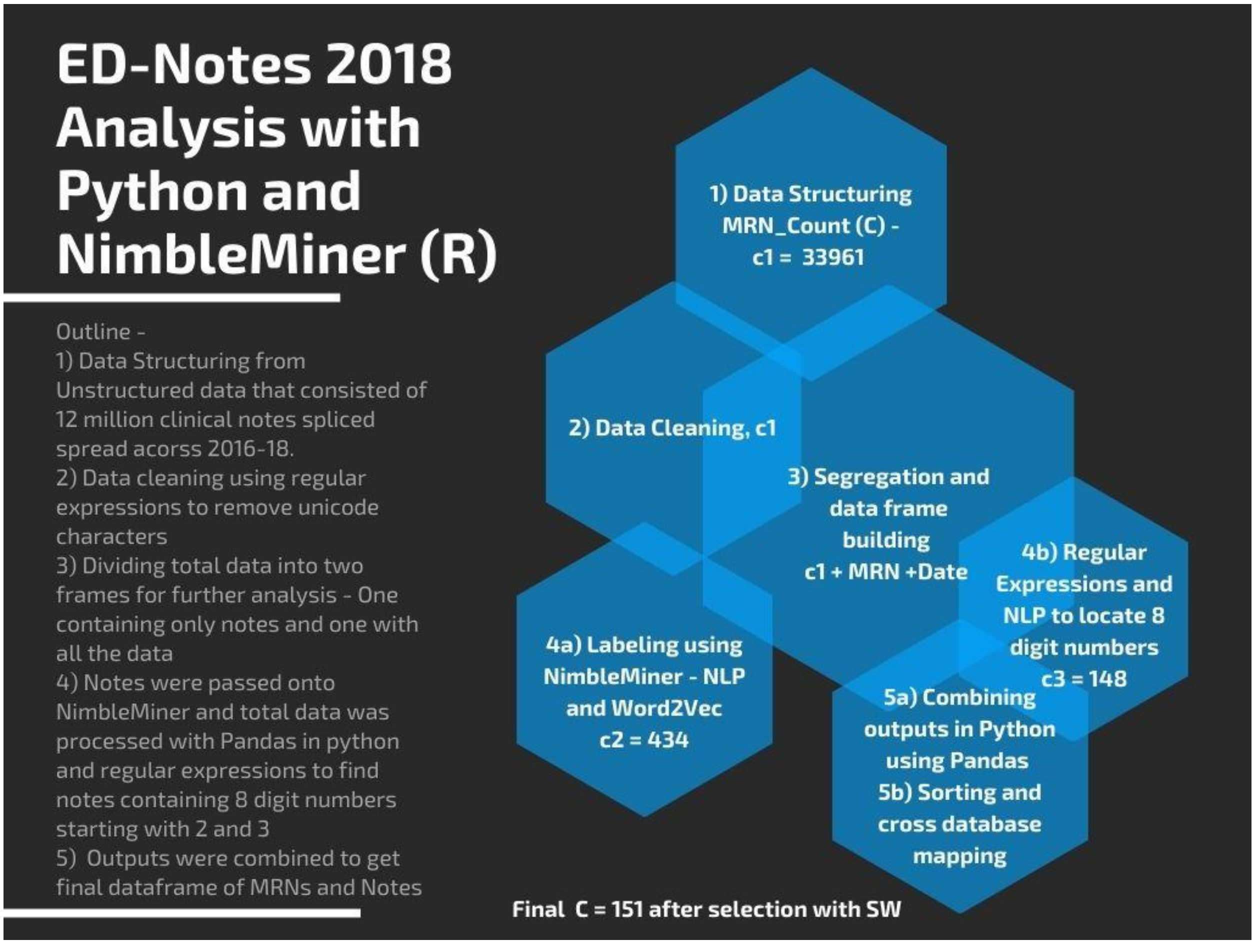
Natural Language Processing Algorithms (NLP-A)(1,2)

(1) Identifying concepts related to Child Abuse and Neglect Reporting using NimbleMiner

(2) Developing an '8-digit' case ID extraction and labeling algorithm

Applying NLP algorithm on a large set* of clinical notes

1. Of then 287757 clinical notes and 33961 unique MRNs, through (NLP-A)(1), we contained at first 434 clinical notes and 254 unique MRN's. After AL manually reviewed the unique MRN's, only **138** unique MRN's remained that displayed abuse and neglect reporting from the ED. (1-2-3-4a)
2. Of the 287757 clinical notes, through the (NLP-A)(2), we contained 148 medical notes and 140 unique MRN's that have 8-digit numbers starting with a two or a 3. After AL manually reviewed the unique MRN's, only **101** unique MRN's remained that displayed abuse and neglect reporting from the ED. (1-2-3-4b)
3. After combing both the 8-digit number and NimbleMiner list of Unique MRN's, we have **151** unique MRN's that represent child abuse and neglect reporting from the ED in 2018.
4. Finally, we combined the list we received from CJC (15 mutually exclusive cases out of 78 total cases) and the list from stage 3 and have **166** cases of child abuse and neglect reporting. (4a+4b => 5a and 5b)



ED-Notes 2018 Analysis with Python and NimbleMiner (R)

Outline -
1) Data Structuring from Unstructured data that consisted of 12 million clinical notes spliced spread acorss 2016-18.
2) Data cleaning using regular expressions to remove unicode characters
3) Dividing total data into two frames for further analysis - One containing only notes and one with all the data
4) Notes were passed onto NimbleMiner and total data was processed with Pandas in python and regular expressions to find notes containing 8 digit numbers starting with 2 and 3
5) Outputs were combined to get final dataframe of MRNs and Notes

1) Data Structuring MRN_Count (C) - c1 = 33961

2) Data Cleaning, c1

3) Segregation and data frame building c1 + MRN +Date

4a) Labeling using NimbleMiner - NLP and Word2Vec c2 = 434

4b) Regular Expressions and NLP to locate 8 digit numbers c3 = 148

5a) Combining outputs in Python using Pandas
5b) Sorting and cross database mapping

Final C = 151 after selection with SW

**The original dataset was of 12 million notes resulting from notes being divided - a combination of unstructured, unsorted 2016-18 data. The DS (Data scientist) - sorted those notes and then divided them into a smaller dataset for accurate and agile analysis. - Agile methodology for extensive data mining through data division.*