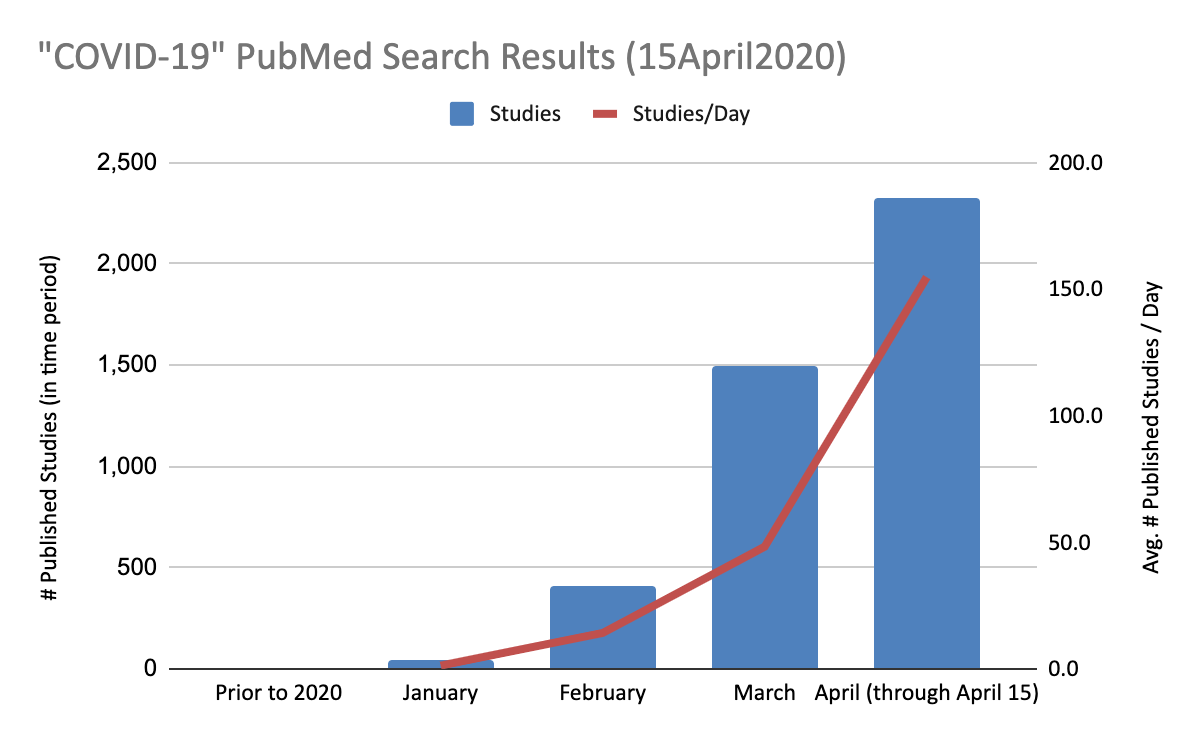
# Problem Statement

COVID-19 is an emerging, rapidly evolving situation. Health care professionals and the public at large are inundated daily with new findings, observations, and studies. A simple “COVID-19” PubMed search on April 15, 2020 quickly frames the issue.



Medical professionals looking to gain perspective from the trusted PubMed repository of studies quickly find that 99.8% of PubMed studies were published in 2020 and the number of daily published studies has rapidly increased beyond an average of 150 studies per day during the first half of April, and the rate is increasing. 262 studies were published on April 14 alone. In the absence of a well defined clinical trial and/or a simple means for gauging the clinical utility of a broad range of studies, how are medical professionals to sort through, understand and enhance their care and treatment and intervention plans from this quickly growing heap of published data?

Goal

Our team brought together by Wonderbench and comprised of data scientists, NLP students, and MedTech professionals aims to use Artificial Intelligence (AI) and Natural Language Processing (NLP) to understand and characterize the clinical utility of therapeutic studies so that the clinical utility and impact of newly published studies can be immediately assessed. In collaboration with the broader community, we look to develop an AI-driven guidance tool that would serve to support medical professionals in quickly highlighting which of the newest studies may be most deserving of their immediate time, attention and consideration toward the goal of rapidly improving the intervention and treatment plan for COVID-19 patients.

Approach

1. Create a corpus of PubMed Studies
   1. We have focused only on studies published in 2020, accessible via the [CORD-19 dataset](https://pages.semanticscholar.org/coronavirus-research).
   2. Extracted all json files of articles and created a unified dataset/table.
   3. Created trivial features such as word count, character count, average word length, number of stopwords, number of special characters, number of numerics.
   4. Cleaned the body text of every article(Removed stop words, lowercased words, removed punctuations and special characters).
   5. -
2. Characterize studies and derive insights from them based on the following approaches:
   1. Approach 1: Analyze the keywords assigned to each article by Pubmed (available via pubmed).
   2. Approach 2: Analyzing based on [MesH (Medical Subject Headings)](https://www.nlm.nih.gov/mesh/meshhome.html). MesH is a controlled semantic vocabulary used by NLM/Pubmed. We plan to use the ‘[MeSH on Demand](https://ii.nlm.nih.gov/Interactive/MeSHonDemand.shtml)’ tool to automatically identify concepts and characterize the articles. For example, COVID-19 is [already a concept](https://meshb.nlm.nih.gov/record/ui?ui=C000657245) in MeSH and we can intersect that with the [Therapeutics concept](https://meshb.nlm.nih.gov/record/ui?ui=D013812) class.
   3. Approach 3: We can also apply other entity extraction tools like [Scispacy](https://allenai.github.io/scispacy/), AWS [Comprehend Medical](https://aws.amazon.com/comprehend/medical/), Google Cloud [NL API](https://cloud.google.com/natural-language).
3. Present the findings and insights to domain experts (Infectious Disease clinicians, researchers) and get feedback on the relative value of the output from the three approaches (and the combination of one/more approaches). Based on that, refine the model and re-run the analysis.
4. Once a useful output is created, we plan to create a continually-updating service that can ingest, extract and analyze content related to Covid-19 related therapeutic research. That service can be utilized by clinicians/researchers in a web or mobile application.

#### 

#### 

#### 

#### We welcome feedback so that we can continue to improve this project.