COLUMBIA UNIVERSITY
IN THE CITY OF NEW YORK

Frederick National Laboratory
for Cancer Research

sponsored by the National Cancer Institute

Master's Research and Projects: FNLCR and Columbia University Partnership

DEPARTMENT OF HEALTH AND HUMAN SERVICES • National Institutes of Health • National Cancer Institute

Frederick National Laboratory is a Federally Funded Research and Development Center operated by Leidos Biomedical Research, Inc., for the National Cancer Institute

Master's Research and Projects: FNLCR and Columbia University Partnership

2:00 PM: Introduction Prof. Michael Robbins

2:10 PM: Opening remarks Dr. Ethan Dmitrovsky

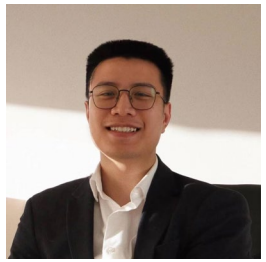
2:20 PM: Student Presentations

- Cloud Deployment, Optimization Strategies for Teaching, Training and Collaborative Reproducible Research
- Survey to Identify Emerging Infectious Disease Datasets for Machine Learning
- Survey to Identify Cancer Datasets for Machine Learning
- Q & A

3:20 PM: Closing remarks Dr. Eric Stahlberg

Project Team

Mahitha Kotipalli

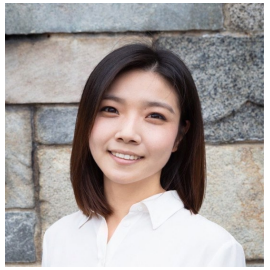


Jim Hu

Niranjana Moleyar

Malin Ortenblad

Kerry Hu



Jie Chen

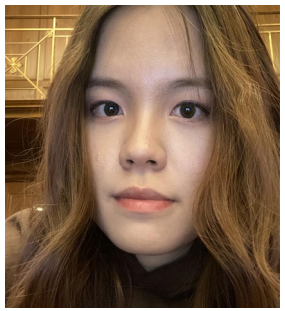
Mengyao He

Om Vaghasia

Panagiotis Misirlis

Jiaxi Zhou

Xinyao Wang



Qinwei Zhang

Yue Hu

Zihui Zhou

Naomi Ohashi, MPA, PMP, ITIL
Biomedical Informatics and Data Science (BIDS)
Frederick National Laboratory for Cancer Research

Ravichandran Sarangan, PhD, PMP
Biomedical Informatics and Data Science (BIDS)
Frederick National Laboratory for Cancer Research

Michael Robbins
Professor
Columbia University

Nicole Soder
TA, Project Manager
Columbia University

COLUMBIA UNIVERSITY
IN THE CITY OF NEW YORK

Frederick National Laboratory
for Cancer Research
sponsored by the National Cancer Institute

Survey to Identify Emerging Infectious Disease Datasets for ML

Team Members: Mengyao He, Om Vaghasia, Panagiotis Misirlis

Sep. 8, 2020

DEPARTMENT OF HEALTH AND HUMAN SERVICES • National Institutes of Health • National Cancer Institute

Frederick National Laboratory is a Federally Funded Research and Development Center operated by Leidos Biomedical Research, Inc., for the National Cancer Institute

Project Team

Mengyao He
Columbia University
MSOR student
Project 2 - COVID-19
Prior experience in data analytics and machine learning. Worked with Q-squared to predict futures contracts price movement by Machine Learning models.

Om Vaghasia
Columbia University
MSMSE Student
Project 2 - COVID-19
Prior experience in data analysis and Natural Language Processing. Worked with International American Supermarkets Corps to optimize their operations.

Panagiotis Misirlis
Columbia University
MSMSE Student
Project 2 - SARS
Prior experience in data analysis, machine learning and deep learning. Worked with a finance firm to create fraud detection tool using ML and with a Non-Profit Organization to create a digital content classifier that is able to detect and count number of Humans in a given picture.

Naomi Ohashi, MPA, PMP, ITIL
Biomedical Informatics and Data Science (BIDS)
Frederick National Laboratory for Cancer Research

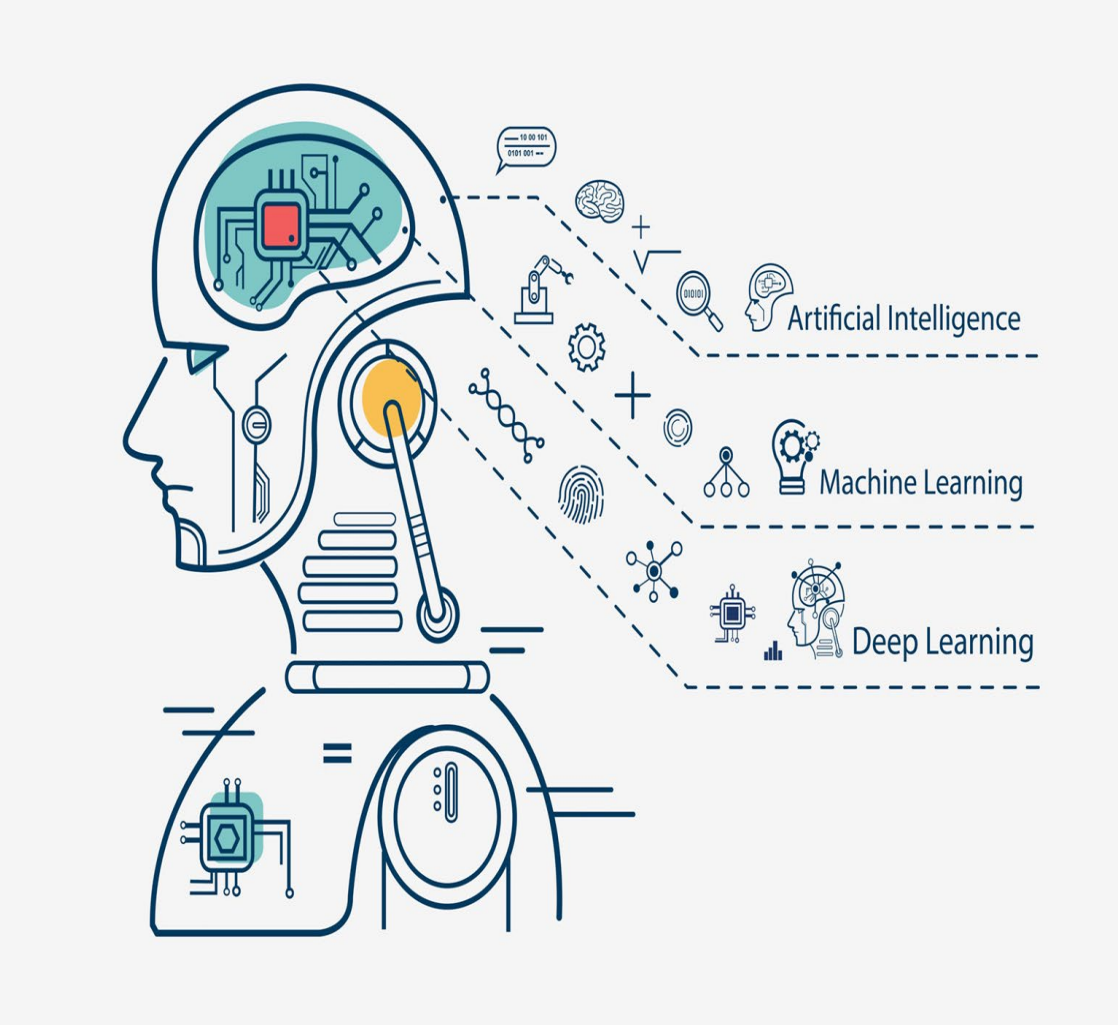
Ravichandran Sarangan, PhD, PMP
Biomedical Informatics and Data Science (BIDS)
Frederick National Laboratory for Cancer Research

Michael Robbins
Professor
Columbia University

Nicole Lynn Soder
TA, Project Manager
Columbia University

Table of Content

Project Overview: Background



Machine-learning (ML) has become a crucial tool in cancer/biomedical research. The important ingredient for a data science project is the data and despite tremendous progress in ML algorithms, data availability remains a major obstacle.

Project Overview: Description

- Used PubMed Advanced Search Builder to ***collect*** and **report** published Machine Learning papers from the area of emerging infectious diseases such as **SARS** (2003) and **COVID-19** (2019)
- For this project, we would prefer the students to focus on ML/AI/Deep-learning modelling efforts that used either **structured** *(Ex. drug SMILES data, DNA/protein, protein-drug binding data)* or **unstructured** *(image data such as X-ray or histopathology etc.)*.
- The student(s) will **document the URL location** of the data that accompanies manuscript(s) and **download the raw (uncleaned) data for preparation and analysis**.
- We **also document datasets that are available only on the web** and not associated with any publication

Project Overview: Goals

- This database will serve as the literature-search starting point for data science projects in NCI/NIH/FNLCR
- The datasets that accompany manuscripts might be ready for modeling and can serve as inputs for proof-of-concept NIH/FNLCR/NCI projects
- Identify any potential trends in the literature in terms of:
  a) programming languages
  b) libraries and packages
  c) specific Machine Learning methodologies (i.e. SVM, Random Forest)

SARS (2003)

Panagiotis Misirlis

Introduction

Literature Review:

- The research for SARS-2003 epidemic was mainly focused on the ***spread*** and ***infection rate*** of the virus in different scenarios/cities.
- Another part of the research papers is focused on ***bio-mapping*** the SARS-COV virus.
- The techniques that have been mostly used are :
  - Neural Networks
  - Decision Trees
  - Stochastic Dynamic Models
- They use ***time-series data*** of ***infections/deaths/recoveries*** that were collected in different cities around the world.
- There are certain limitations with SARS literature review and the most significant one is ***time elapsed*** since that epidemic happened. As we are going to see also later on this has a significant impact on the type of techniques that were used.

Methods to Conduct Literature Reviews

PubMed Keywords

- Pubmed search engine:
  - https://pubmed.ncbi.nlm.nih.gov/
- SARS:
  - SARS 2003 Machine Learning
  - SARS 2003 Forecasting
  - SARS 2003 Prediction
  - SARS 2003 Classification

Other Methods

- Identified other useful publications from footnotes and references.
- Kaggle
- Github Search

Results: Major Groups in SARS Studies

Forecasting & Prediction

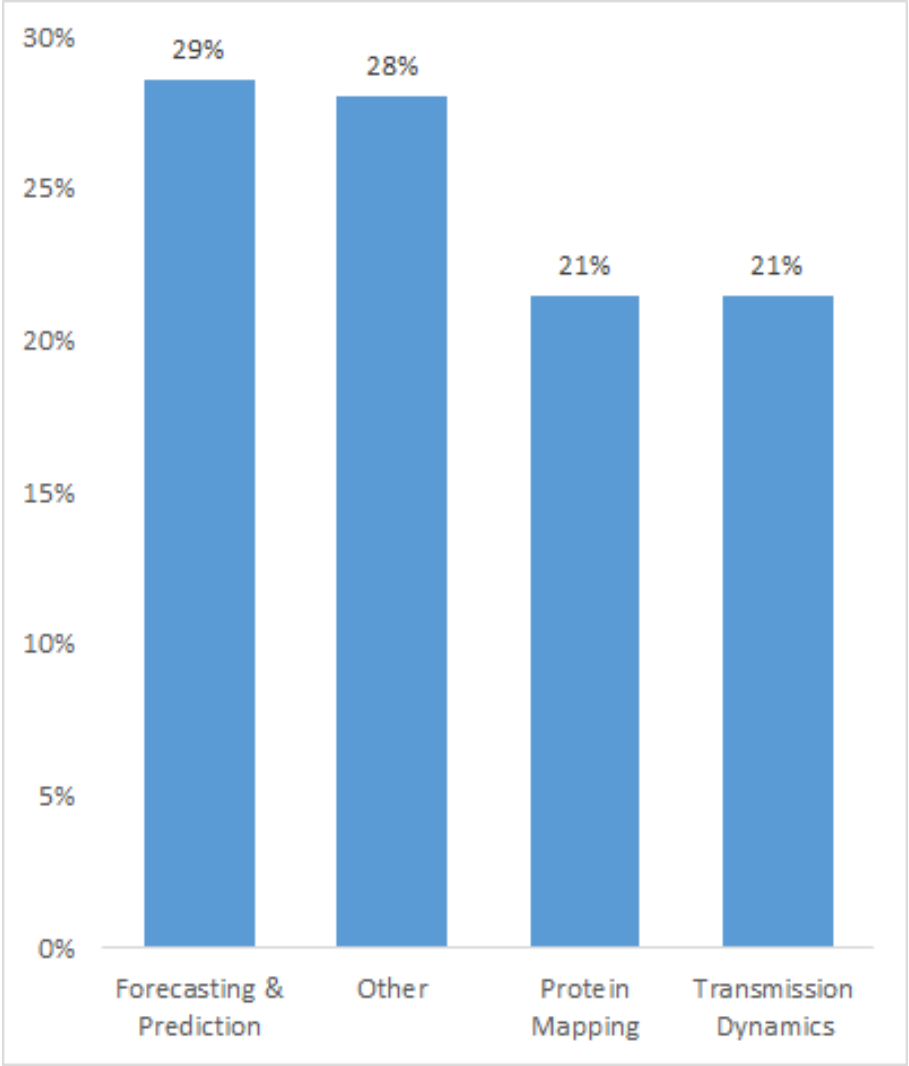
- Forecast the trajectory of the outbreak and quantify the risk of death.
- Predict the transmission rates of the virus under different circumstances

Protein Mapping

- Prediction Rule generation for SARS-CoV protease cleavage sites.
- Mining SARS-CoV protease cleavage data

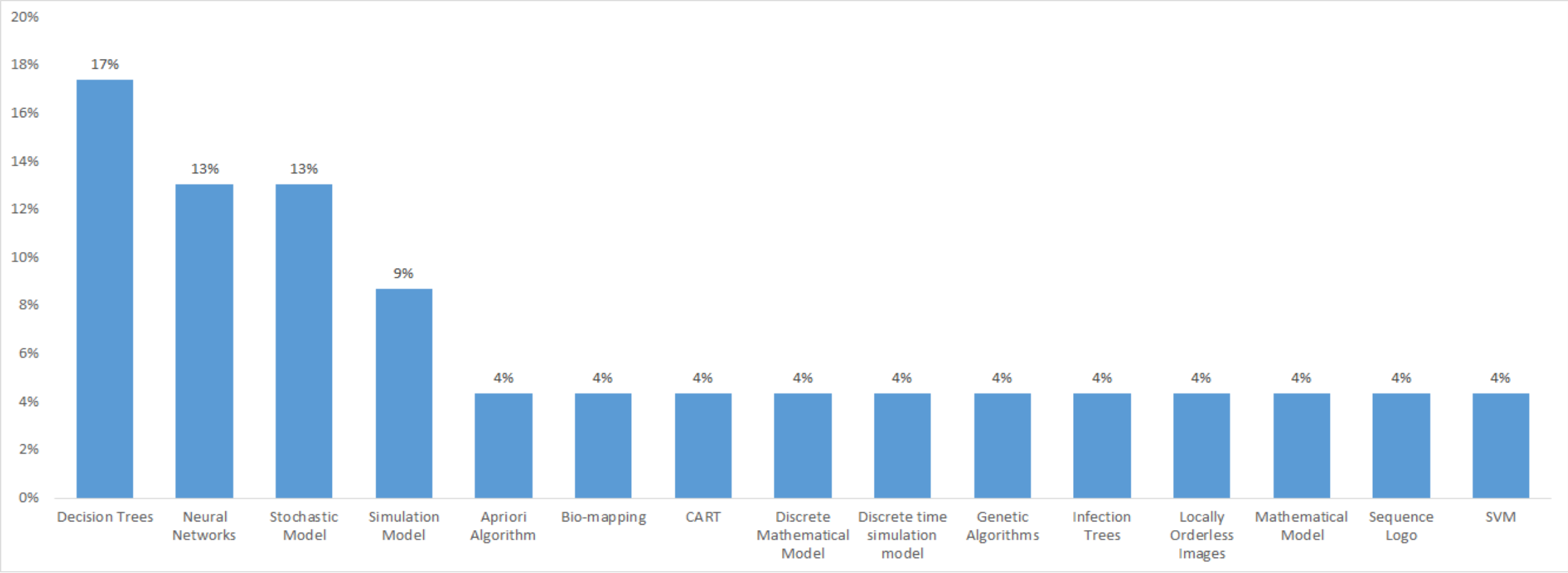
Transmission Dynamics

- Transmission Models for SARS-COV virus
- Superspreading and the effect of individual variation on disease emergence



Results: Popular Softwares and ML Models

Popular techniques and algorithms used for SARS studies



Results: Popular Softwares and Packages

Popular Softwares and Packages:

- Only one of the papers mentions a coding language, which is Java and C. All of the other papers don't mention how the models were created (i.e. using a specific program or a programming language) and no links were found to any type of code/repositories.We believe that's due to the time period of the virus as before mentioned. Majority of the papers that were found date back more than 15 years.

Quality of papers and dataset:

- Large part of the papers don't include their datasets. The ones that do include links to datasets, the links are out of date and aren't reachable.
- External Datasets were found to fill the gap of data. The data that has been found is 1) Case/Deaths Time Histories for many different parts of the world and 2) X-Ray images from SARS patients.

Results: Datasets from Machine Learning Publications

	Dataset	Link
1	Table with global data for the SARS epidemic	https://github.com/WL-Biol185-ShinyProjects/sars-project/blob/master/SARS%20data.xlsx
2	The file contains day by day no. from March to July 2003 across the world.	https://www.kaggle.com/imdevskp/comments
3	Final summary data from across the world	https://www.kaggle.com/imdevskp/comments
4	Chest X Ray dataset for SARS	https://github.com/mlmed/torchxrayvision

COVID-19

Mengyao He, Om Vaghasia

Introduction

Literature Review:

- Researchers are using Machine Learning techniques to apply to problems related to COVID-19, from ***diagnosing early symptoms*** using Chest X-rays to internet search queries to ***estimate the effect on mental health of social distancing***.
- Most of the research is focused on classification and prediction.
- The techniques that have been mostly used are :
  - K-means, K-nearest neighbors
  - Random Forest and tree-based models in general
  - Neural Networks and Deep Learning
- They mostly use ***time-series data*** of ***infections/deaths/recoveries, chest X-rays and CT scan,*** that were collected in different cities around the world.
- Heavily subsidized research topic, leading to a vast amount of research papers and datasets.

Methods to Conduct Literature Reviews

PubMed Keywords

- Pubmed search engine:
  - https://pubmed.ncbi.nlm.nih.gov/
- COVID-19:
  - COVID19 Machine Learning
  - SARS-CoV-2 Machine Learning
- Results:
  - 59 publications
  - 34 publications datasets

Other Methods

- Harvard Dataverse
- Kaggle
- Github repositories
- Links provided within other sections of the papers

Results: Major Groups in ML-based Covid-19 Studies

Screening

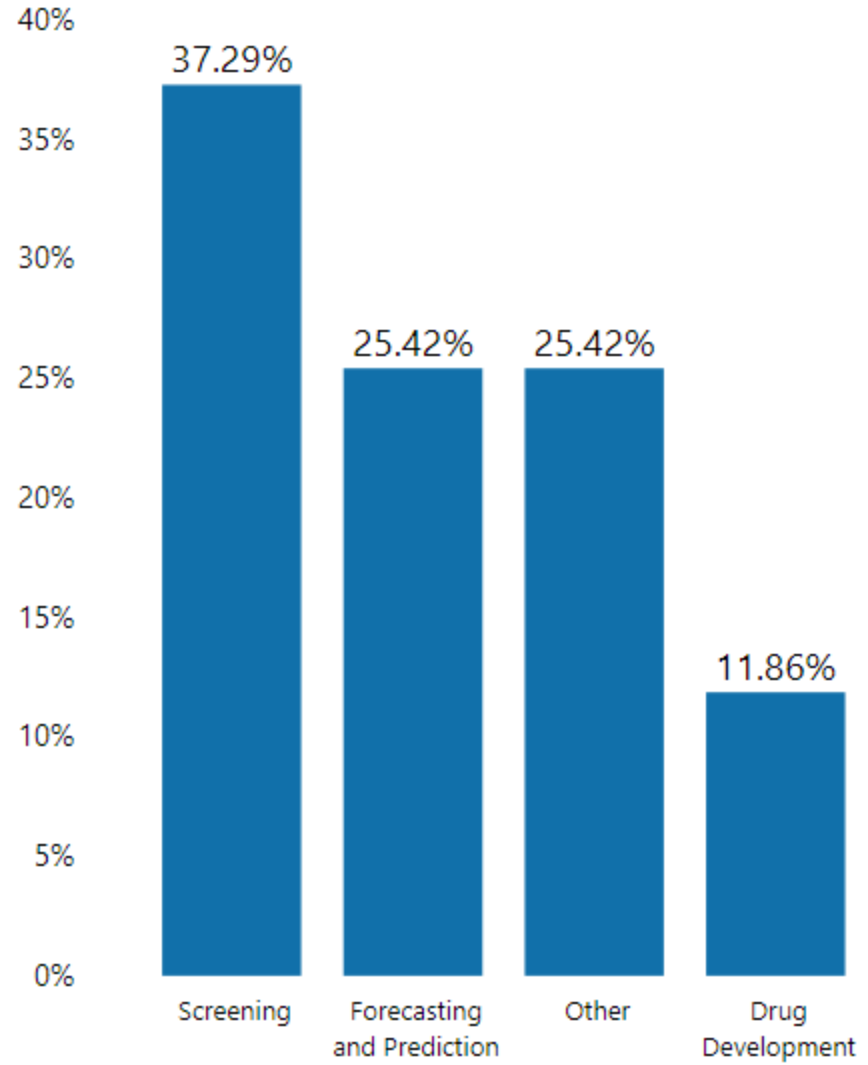
- Extract radiological features for timely and accurate COVID-19 diagnosis from CT images and X rays.
- Distinguish COVID-19 from community acquired Pneumonia and other lung diseases.

Forecasting & Prediction

- Forecast the trajectory of the outbreak and quantify the risk of death.
- Show feasibility and accuracy for predicting hospital stay in COVID-19 patients
- Protein structure predictions

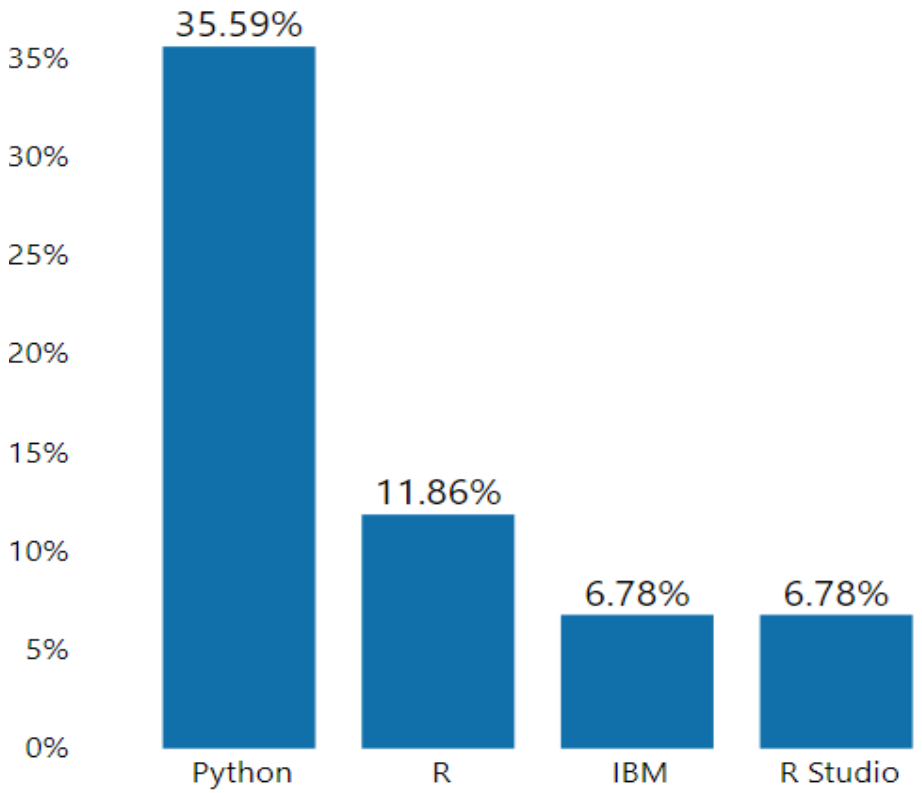
Drug Development

- Computational biology and medicines perspective
- Generate novel drug compounds

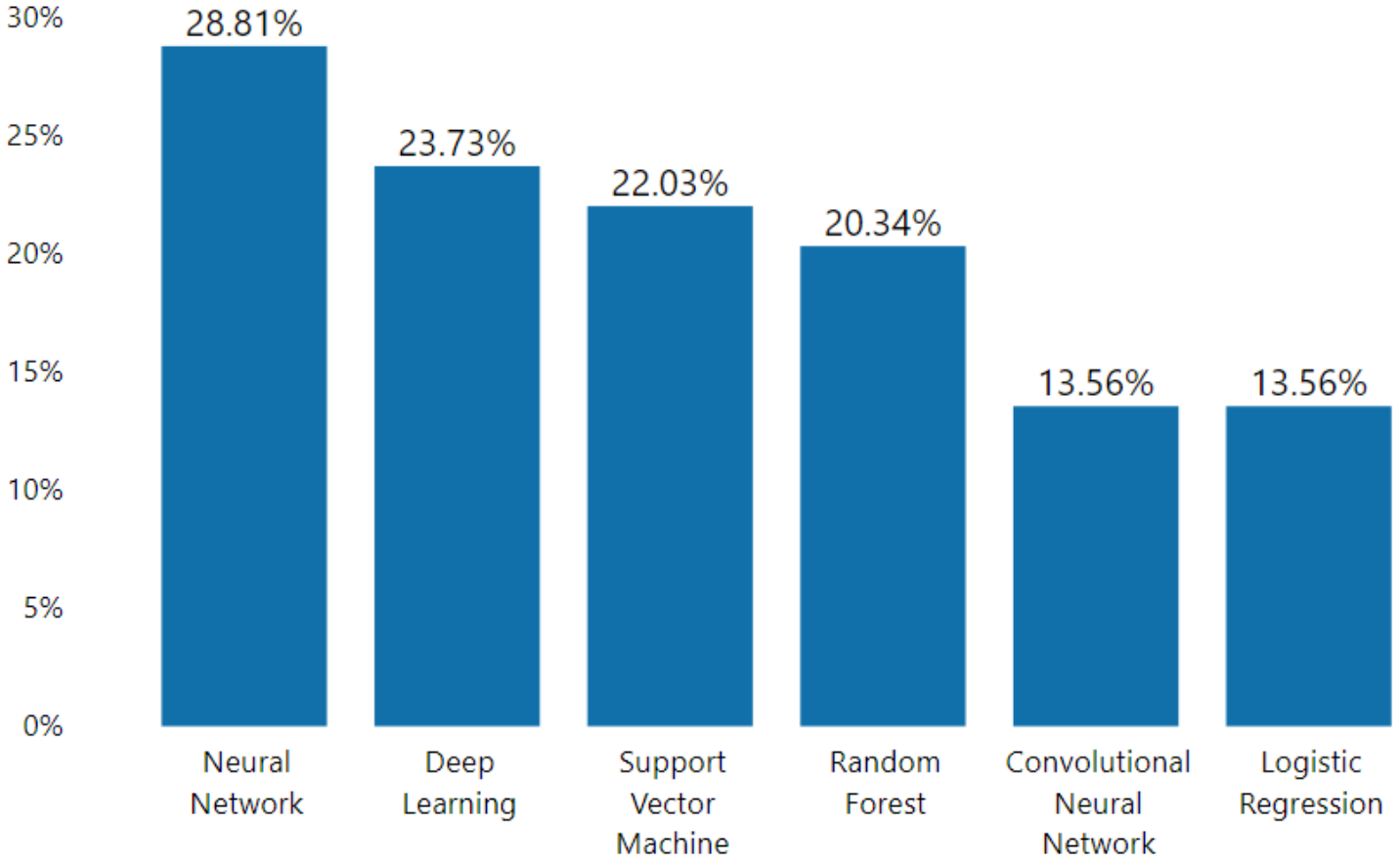


Results: Popular Softwares and ML Models

Popular softwares used for ML-based studies for COVID-19



Popular ML models and algorithms used for COVID-19 studies



Results: Datasets from COVID-19 Machine Learning Publications

	Dataset	Link		Dataset	Link
1	Pulmonary Chest CT Scans	https://github.com/bkong999/COVNet	10	Baidu migration	http://qianxi.baidu.com/
2	COVID-19 Affected Cases	https://www.kaggle.com/sudalairajkumar/novel-corona-virus-2019-dataset	11	Real-time cases of COVID-19	https://news.qq.com/zt2020/page/feiyan.htm?ADTAG=area
3	COVID-Chestxray-Dataset	https://github.com/ieee8023/covid-chestxray-dataset	12	Situation report 2020	http://www.nhc.gov.cn/xcs/yqtb/list_gzbd.shtml
4	CSSEGISandData	https://github.com/CSSEGISandData/COVID-19	13	Combatting SARS 2003	http://news.sohu.com/57/26/subject206252657.shtml
5	Protein Sequences	https://bigd.big.ac.cn/ncov	14	Dataset of SARS-CoV-2 Genome	https://data.mendeley.com/datasets/nvk5bf3m2f/1
6	Proteome Data	https://www.iprox.org/page/ProjectFileList.html?projectId=IPX0002106000	15	COVID-Classifier	https://github.com/abzargar/COVID-Classifier/tree/master/dataset
7	MLDSP-GUI	https://sourceforge.net/projects/mldsp-gui/files/COVID19Dataset/	16	NYT Data Collection	https://github.com/mihirpsu/covid_19/tree/master/Data_Collection
8	COVID-19 Cases Dataset	https://www.kaggle.com/imdevskp/corona-virus-report	17	COVID-19 Brodinlab	https://ki.app.box.com/s/sby0jesyu23a65cbgv51vpbzqjdmipr1
9	COVID-19 Global Wealth	https://www.kaggle.com/winterpierre91/covid19-global-weather-data	18	UCSDAI4H COVID-19 CT	https://github.com/UCSD-AI4H/COVID-CT

Results: Datasets from COVID-19 Machine Learning Publications

	Dataset	Link		Dataset	Link
19	Confirmed COVID-19 cases	https://doi.org/10.6084/m9.figshare.12030363.v1	27	GCCR001	https://osf.io/a3vkw/
20	Multimedia component	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7186211/bin/mmc1.csv	28	100 axial CT images	http://www.salhospital.com/
21	Health Report from China CDC	https://github.com/midas-network/COVID-19/tree/master/data/cases/china	29	Public Chest X-Ray (CXR) datasets	https://www.isi.uu.nl/Research/Databases/SCR/download.php
22	CoronaHack CXR Dataset	https://www.kaggle.com/praveengovi/coronahack-chest-xraydataset	30	Montgomery County X-ray Set	https://academictorrents.com/details/ac786f74878a5775c81d490b23842fd4736bfe33
23	CXR Matedata	https://github.com/jongcye/Deep-Learning-COVID-19-on-CXR-using-Limited-Training-Data-Sets/blob/master/metadata.xls	31	JSRT Database	http://db.jsrt.or.jp/eng.php
24	Confirmed COVID-19 cases	https://doi.org/10.6084/m9.figshare.12030363.v1	32	GCCR001	https://osf.io/a3vkw/
25	Multimedia component	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7186211/bin/mmc1.csv	33	100 axial CT images	http://www.salhospital.com/
26	COVID Chestray Dataset	https://github.com/ieee8023/covid-chestxray-dataset	34	Serial chest radiographs graphs	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7135076/figure/fig1/

SARS (2003) and COVID-19 Conclusions

SARS (2003)

- Most of the methodologies that were used in the literature are outdated. The most common techniques found are Neural Networks, Decision Trees and Stochastic Models.
- In terms of data all of the articles weren't useful in providing data. External Datasets were found.
- Majority of papers date back more than 15 years, opportunity for researchers to use newer ML techniques on the SARS epidemic.

COVID-19

- The most common techniques found are Neural Networks, Deep Learning, and SVM.
- K-Nearest Neighbors is observed to be a preferred method for clustering.
- Size of datasets for studies on Screening seems to be an issue, particularly for Deep Learning and Convolutional Neural Network algorithms. Transfer Learning can be explored as a potential solution.

Future Work

SARS (2003)

Researchers could use the SARS datasets that have been found and the methods from the COVID-19 Machine Learning papers to implement more up-to-date techniques and see how they compare to the ones that were used 15 years ago.

COVID-19

Publications on ML-based COVID-19 studies are being conducted and uploaded continuously which can be studied and added to the current database.
Trends regarding the type of studies being conducted - Screening, Forecasting and Prediction, and Drug Development - can be tracked to identify some of the needs.

References

- Lalmuanawma S, Hussain J, Chhakchhuak L. Applications of machine learning and artificial intelligence for Covid-19 (SARS-CoV-2) pandemic: A review. *Chaos Solitons Fractals*. 2020;139:110059. doi:10.1016/j.chaos.2020.110059

- Kumar A, Gupta PK, Srivastava A. A review of modern technologies for tackling COVID-19 pandemic. *Diabetes Metab Syndr*. 2020;14(4):569-573. doi:10.1016/j.dsx.2020.05.008

Thank You!

Thank you for this opportunity, especially to Ravi and Naomi. We learned about FNL and about collaborating to solve real machine learning problems in the medical science. We've learned a lot and we hope to be able to collaborate with FNL again.

We have identified ***70 publications and 38 publicly available datasets*** for machine learning studies in the SARS space.

Any questions?