Covid-19 Severe Outcome Risk Prediction

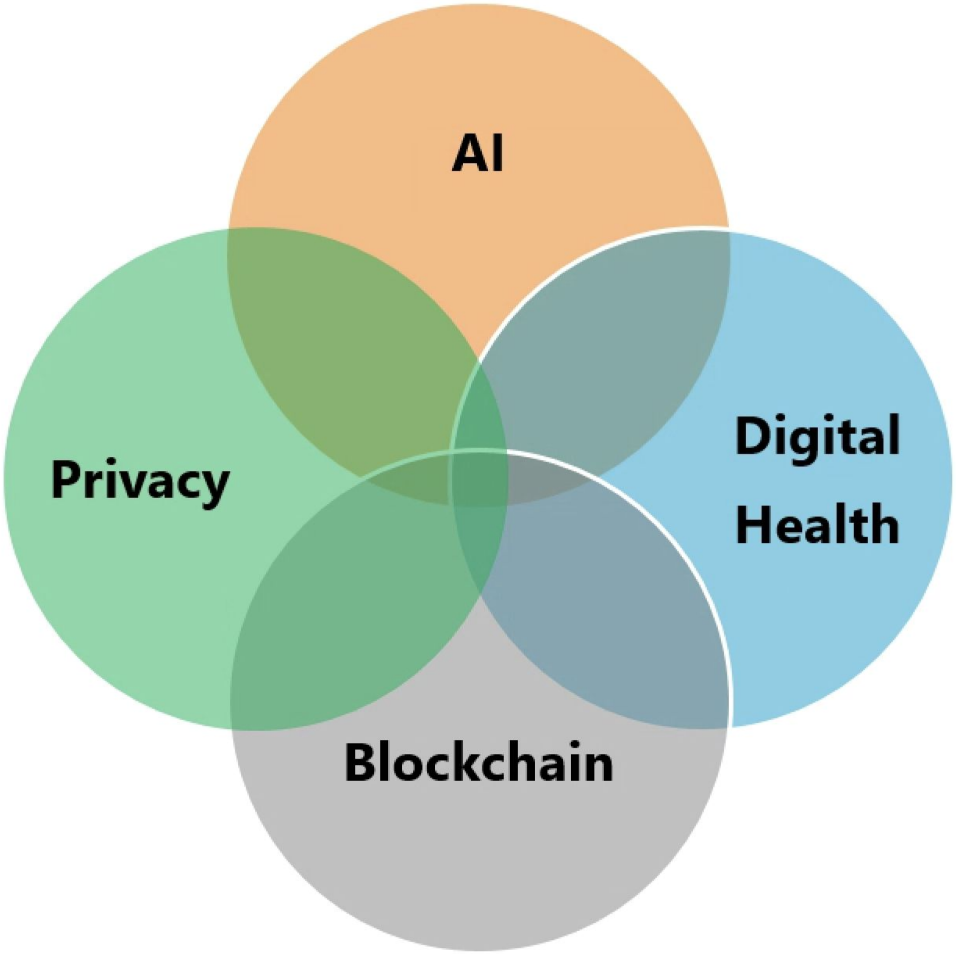
Private Machine Learning on Medical Records & Social Data

Changrong Ji
Dr. Mahesh Shukla
Dr. David Patton
Dr. Xue Yang
Dr. Xingguo Zhang
Antonio Linari
Premdutt Gaur
Vance Degen
A3.AI

Topics

- About Us
- Aims
- Data
- Approach
- Early Findings
- Future Work

Nonprofit Applied R&D



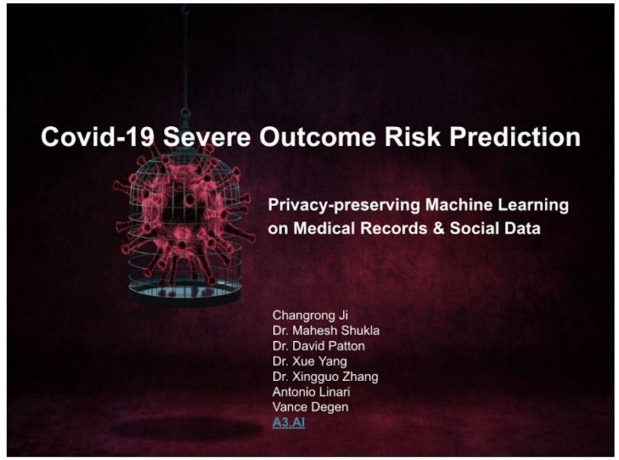
We are:

Data Scientists
Physicians
Engineers
Advocates
Researchers
Privacy Specialists
Game Developer
Venture & Social Capitalist

[A3.AI](A3.AI)

Projects

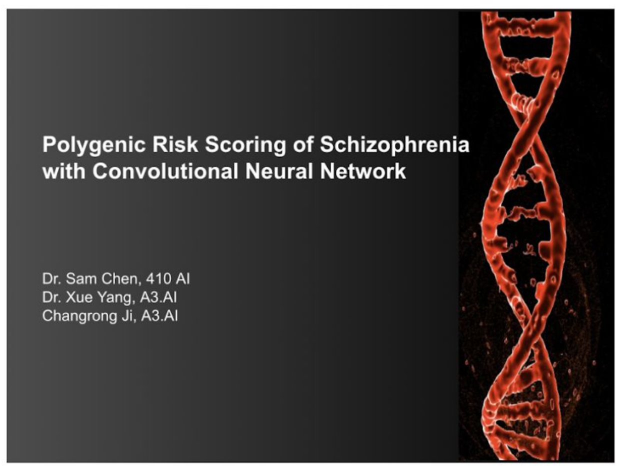
Private & Secure AI



Health Analytics

Drug Safety & Repurposing

Blockchain & Decentralized AI



Genomics

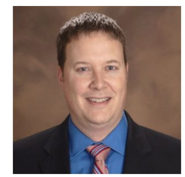
Reinforcement Learning

A3.Ai

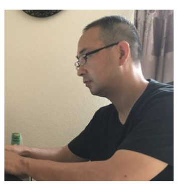
COVID-19 Project Team


Changrong Ji

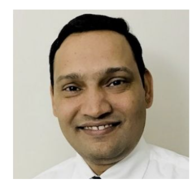

Dr. Mahesh Shukla


Dr. David Patton


Dr. Xue Yang


Dr. Xingguo Zhang


Antonio Linari


Premdutt Gaur


Vance Degen

COVID-19 RESEARCH DATABASE

The data, technology, and services used in the generation of these research findings were generously supplied pro bono by the COVID-19 Research Database partners
https://covid19researchdatabase.org/

Aims

1. COVID-19 Severe Outcome Risk Prediction
2. Social Determinants of Health and Risk Factors of COVID-19
3. Privacy-preserving Machine Learning
4. Building Clinical Concept Embeddings under Computing Resource Constraints

Multiple research aims are addressed in detail in the following working papers respectively as of 09/2020. Future versions will be published as the research progresses:

1. Early findings from machine learning baseline models to predict an individual's risk of hospitalization if infected with COVID-19, based on medical claims and EHR, respectively: Predicting Risk of Hospitalization among COVID-19 Patients, Mahesh Shukla et al.

2. Social Determinants of Health for COVID-19 Diagnosed Patients, David Patton et al.

3. The first in a series of private AI techniques which will enable the release of the AI models built with personal level data while preserving privacy: Privacy-Preserving Machine Learning Techniques 2020, Changrong Ji et al.

4. Clinical Concept embedding is a feature engineering technique to enhance the accuracy of AI models. To build embeddings from large claims data typically requires high computing power. We use a novel approach to efficiently build embeddings under resource constraints in the COVID-19 Research DB environment : Building Clinical Concept Embeddings under Computing Resource Constraints, Antonio Liniari et al.

Data

As of 08/21/2020, with new data added with 1 week delay

- **Claims**
  - 98 million patients 7 years of medical claims history of over 3 billion claim lines
    - Key attributes: ICD diagnosis codes, CPT procedure codes
  - 200,000+ COVID patients
- **Electronic Health Record (Outpatient)**
  - 36 million patient's outpatient EHR records;
    - Diagnoses, Procedures; Encounters, Medications; Allergy, Social History, etc.
  - 16,000 confirmed COVID patients , and 75,000 possible COVID patients
- **Social - Claims - Death linked**
  - 242 million people's Social Data
    - People (demographics, finance, credit, housing, jobs, lifestyle)
    - Behaviors (interests, purchasing, social network activity, charitable giving, health lifestyle)
    - Predictors (motivator, travel, auto, in-market, and economic stats)
  - Death Registry of 80% of US population
    - Died_in_2020
  - 95,000 COVID patients

A3.Ai

Attributions to Data Providers

- AnalyticsIQ

  AnalyticsIQ is s a leading predictive data and analytics innovator that leverages a blend of publicly available data and custom algorithms informed by cognitive psychology concepts to describe consumers across three areas - People, Behaviors, and Predictors. Headquartered in Atlanta and recently named one of Georgia's Top 10 most innovative companies, AnalyticsIQ's team of data analysts, scientists, and cognitive psychologists have over 100 years of collective analytical experience and expertise.

- Health Jump

  Electronic Health Record data including diagnosis, procedures, labs, vitals, medications and histories sourced from participating members of the Healthjump network.

- De-identified claims data was contributed by a claims clearinghouse.

Machine Learning for Clinical Prognosis

Aim 1

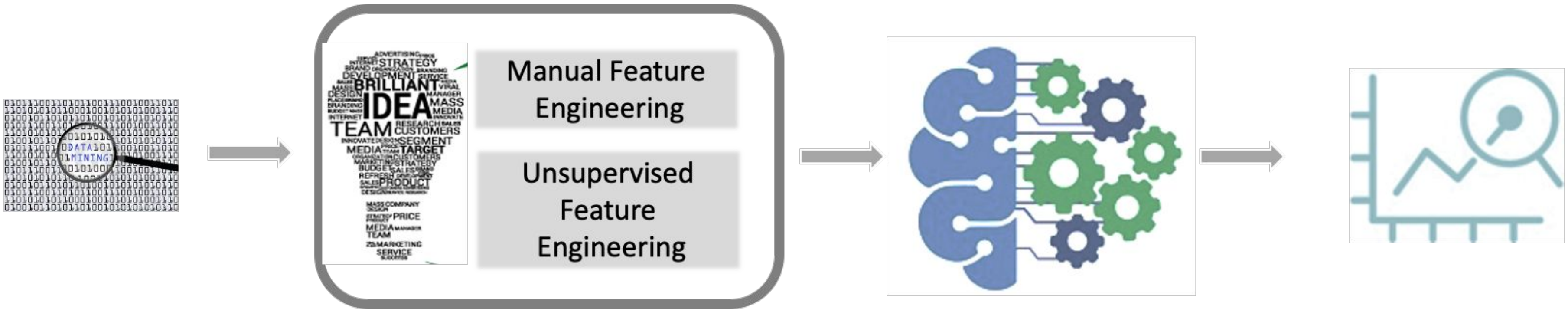
Create machine learning models to predict a patient's risk of severe clinical outcomes if infected with COVID-19.

- Hospitalization
- ICU
- Intubation
- Ventilation
- ECMO (heart-lung bypass)
- Death, etc

These personalized risk scores and associated risk factors analysis can

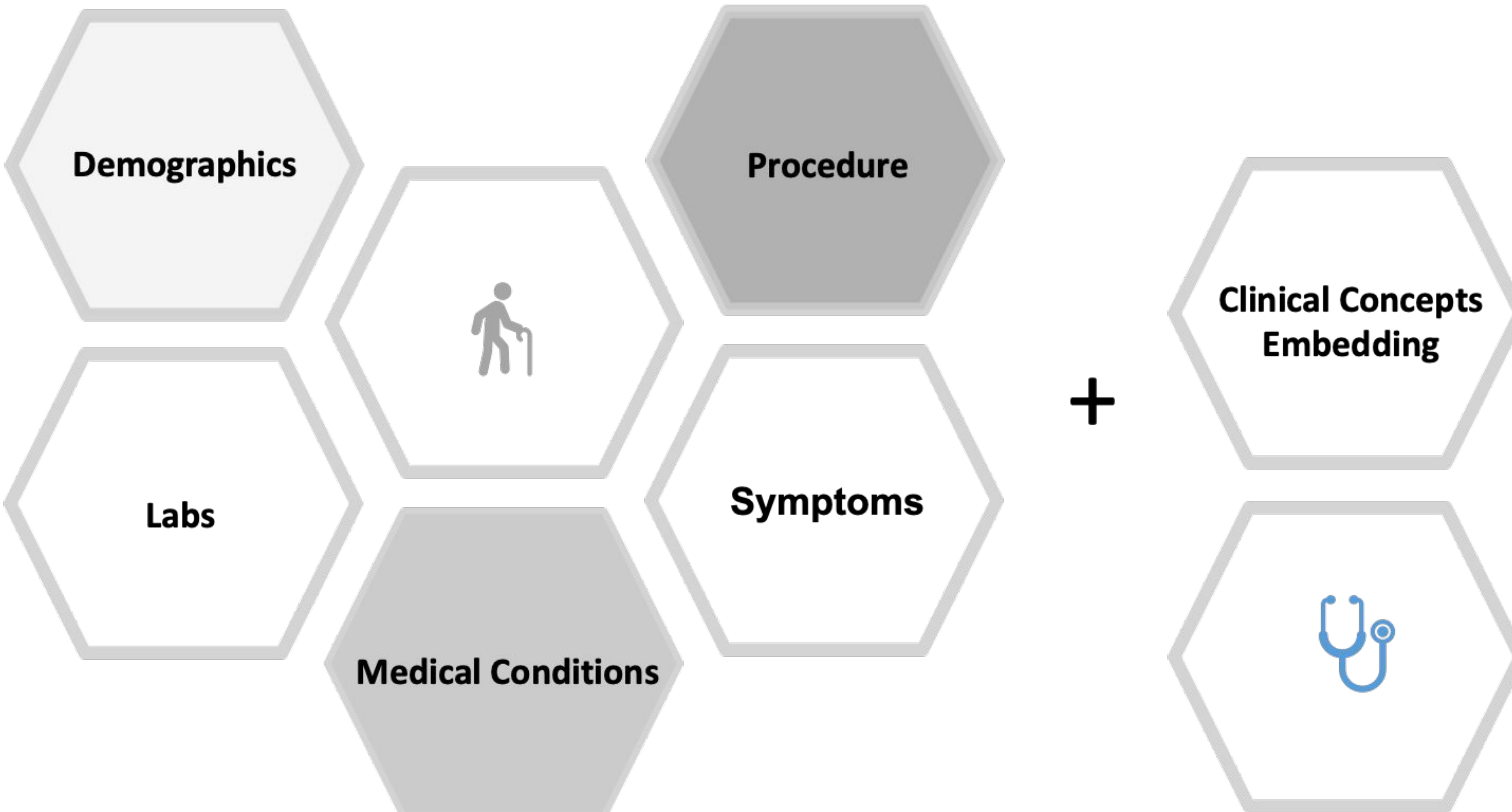
- Help citizens make informed work and lifestyle choices
- Augment clinical prognosis by physicians
- Help health care organizations coordinate care and optimize resources
- Help public health agencies with planning, responding and reopening.

ML Model Development



Raw data **Feature Engineering** **Machine learning Algorithm** **Predictive Models**

Feature Engineering



Embedding in NLP

- A great way to represent sparse, high-dimensional data in NLP

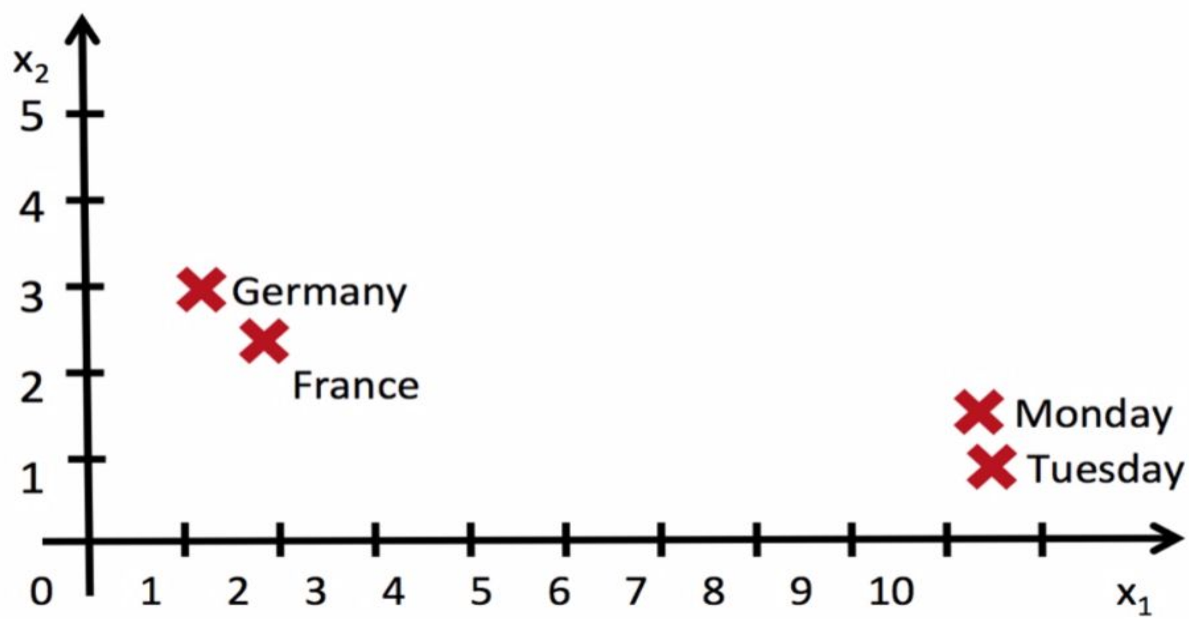


Figure (edited) from Bengio, "Representation Learning and Deep Learning", July, 2012, UCLA

Lower-dimensional space:

- Words of similar meaning are located near each other in the embedded vector
- Relative location of two words in the space could encode a meaningful relationship

Clinical Concepts Embedding

Low dimensional continuous representation of high dimensional discrete data (~300,000 distinct codes)**.**

One type of pretrained model used for transfer learning

Clinical Concepts Embedding

- Low-dimensional vector representations of medical concepts

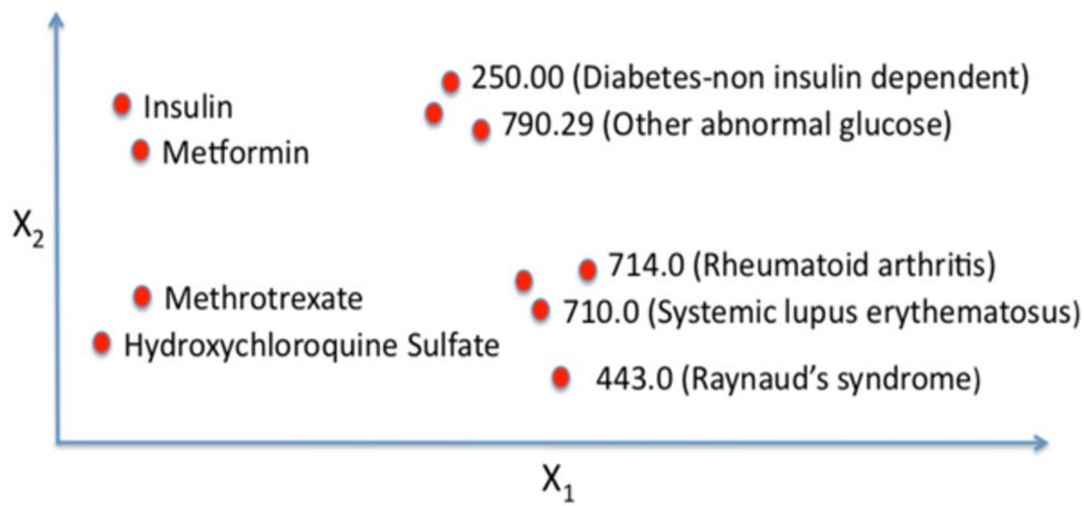


Figure 1: Illustration a low-dimensional representation (in this case, 2 dimensions) of medical concepts. Similar concepts are close to each other in Euclidean space.

Chio. Y, AMIA Jt Summits Transl Sci Proc. 2016 Jul 20;2016:41-50.

Data Labeling

Hospitalization	15.7%
ICU	6.5%
Ventilation	0.4%
ECMO (heart-lung bypass)	0.01%
Death, etc	3.1%

Top 20 Procedures for COVID Patients

CLM + CLM_INST+SVC+SVC_INST	COVID cases 98,182	7/19/2020
PROCEDURE_CODE	Number of patients	Procedure description
U0003	13915	COVID testing using high-throughput technology
99213	11707	Established patient office visit
99285	10967	ER visit for E&M
86769	9432	SARS Cov-2 Antibody test
99233	8666	Initial hospital inpatient care
93010	7938	ECG
	7498	NULL
99223	7390	Initial hospital inpatient care
99291	7332	Critical care, first hour
99232	7254	Subsequent hospital care
71045	6456	Diagnostic imaging chest
99284	4900	ER visit for E&M
87635	4463	SARS Cov-2 DNA RNA test
99283	4436	ER visit for E&M
99309	4393	Subsequent nursing facility care
99212	4271	Office visit E&M
85025	4098	Lab automated diff wbc count
99214	3619	Office visit E&M
80053	3296	Compr metabolic panel
36415	3243	Collection of venous blood
99308	2737	Subsequent nursing facility care

A3.Ai

Top 20 Co-occurring Diagnosis with COVID-19

Diagnosis code	# patients	Diagnosis description
J1289	12321	Other viral pneumonia
I10	10671	Essential hypertension
J9601	8078	ARF with hypoxia
J189	7885	Pneumonia unspecified organism
R05	7118	Cough
R0602	6165	Shortness of breath
E119	5375	Diabetes Mellitus
R509	4749	Fever, unspec
R0902	4325	Hypoxemia
N179	4073	AKF, unspeci
E785	3271	Hyperlipidemia, unsp
A419	3185	Sepsis, unspecified org
Z20828	2832	Contact
N390	2652	UTI, site not sp
M6281	2283	Muscle weakness gene
J988	2235	Other sp resp disorders
D649	2228	Anemia, unsp
R531	2063	weakness

Baseline Prediction with Claims Data

Goal: Prediction of hospitalization for COVID-19 patient

Data:

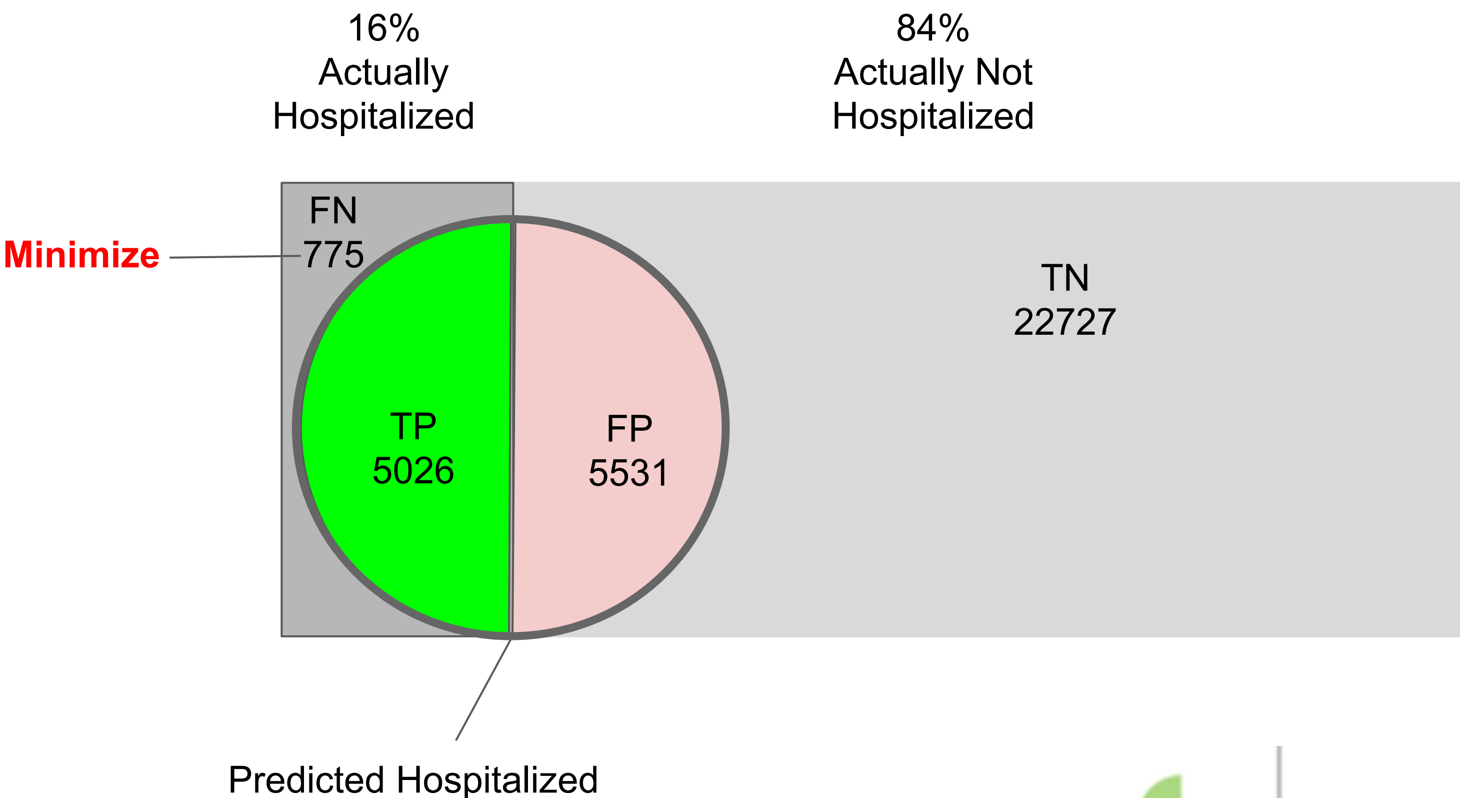
- Hand-crafted features (~100)
- Total patients with COVID-19 infection: 170,241
- Hospitalized patients: 17% of above
- Train/Test/Validation set split: 60/20/20

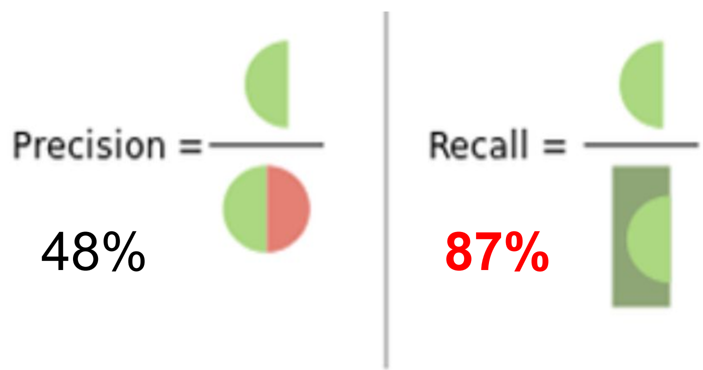
Model:

- Random Forest Classifier
- Balanced weights
- 400 estimators with a max_depth of 20

A3.Ai

Precision & Recall Refresher

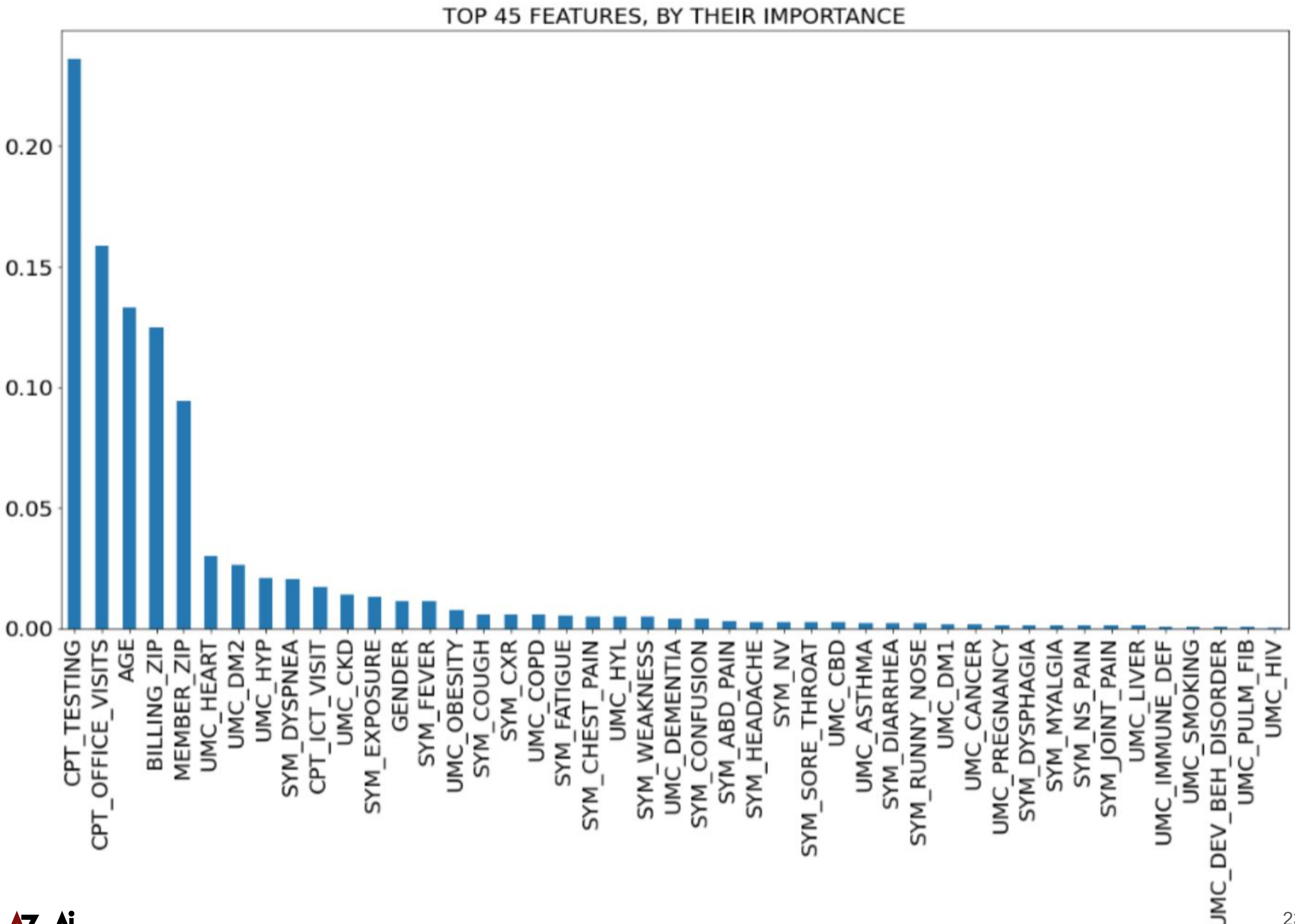




Hospitalization Results

Classification Report:

	precision	recall	f1-score	support
0	0.97	0.80	0.88	28248
1	0.48	0.87	0.61	5801
accuracy			0.81	34049
macro avg	0.72	0.84	0.75	34049
weighted avg	0.88	0.81	0.83	34049


TOP 45 FEATURES, BY THEIR IMPORTANCE
0.20
0.15
0.10
0.05
0.00
CPT_TESTING
CPT_OFFICE_VISITS
AGE
BILLING_ZIP
MEMBER_ZIP
UMC_HEART
UMC_DM2
UMC_HYP
SYM_DYSPNEA
CPT_ICT_VISIT
UMC_CKD
SYM_EXPOSURE
GENDER
SYM_FEVER
UMC_OBESITY
SYM_COUGH
SYM_CXR
UMC_COPD
SYM_FATIGUE
SYM_CHEST_PAIN
UMC_HYL
SYM_WEAKNESS
UMC_DEMENTIA
SYM_CONFUSION
SYM_ABD_PAIN
SYM_HEADACHE
SYM_NV
SYM_SORE_THROAT
UMC_CBD
UMC_ASTHMA
SYM_DIARRHEA
SYM_RUNNY_NOSE
UMC_DM1
UMC_CANCER
UMC_PREGNANCY
SYM_DYSPHAGIA
SYM_MYALGIA
SYM_NS_PAIN
SYM_JOINT_PAIN
UMC_LIVER
UMC_IMMUNE_DEF
UMC_SMOKING
UMC_DEV_BEH_DISORDER
UMC_PULM_FIB
UMC_HIV

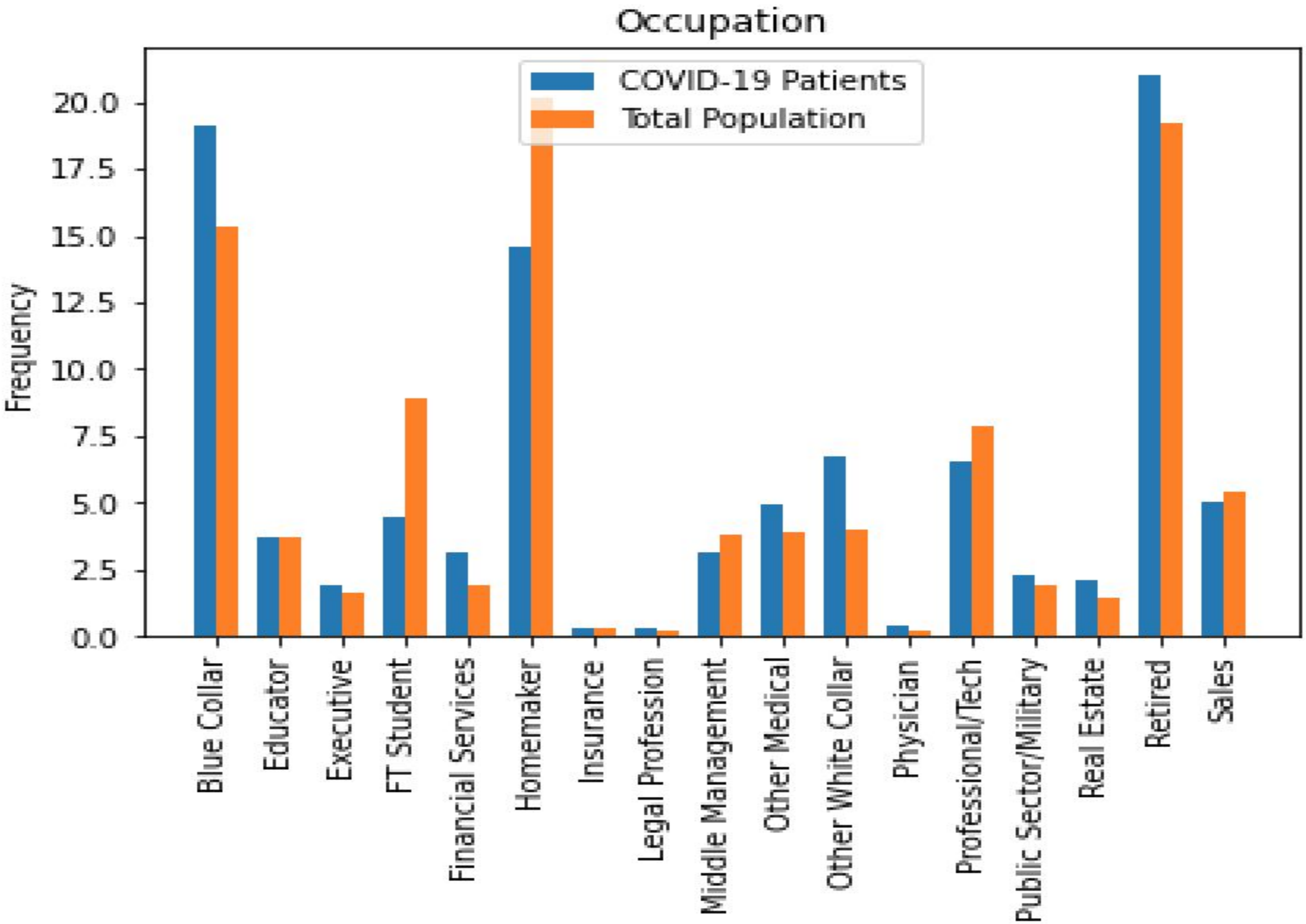
Social Determinants and Risk Factors

About 90 attributes of social data from Analytics IQ are available for over 34 million patients. Over 95,000 are COVID-19 patients.

We examined:

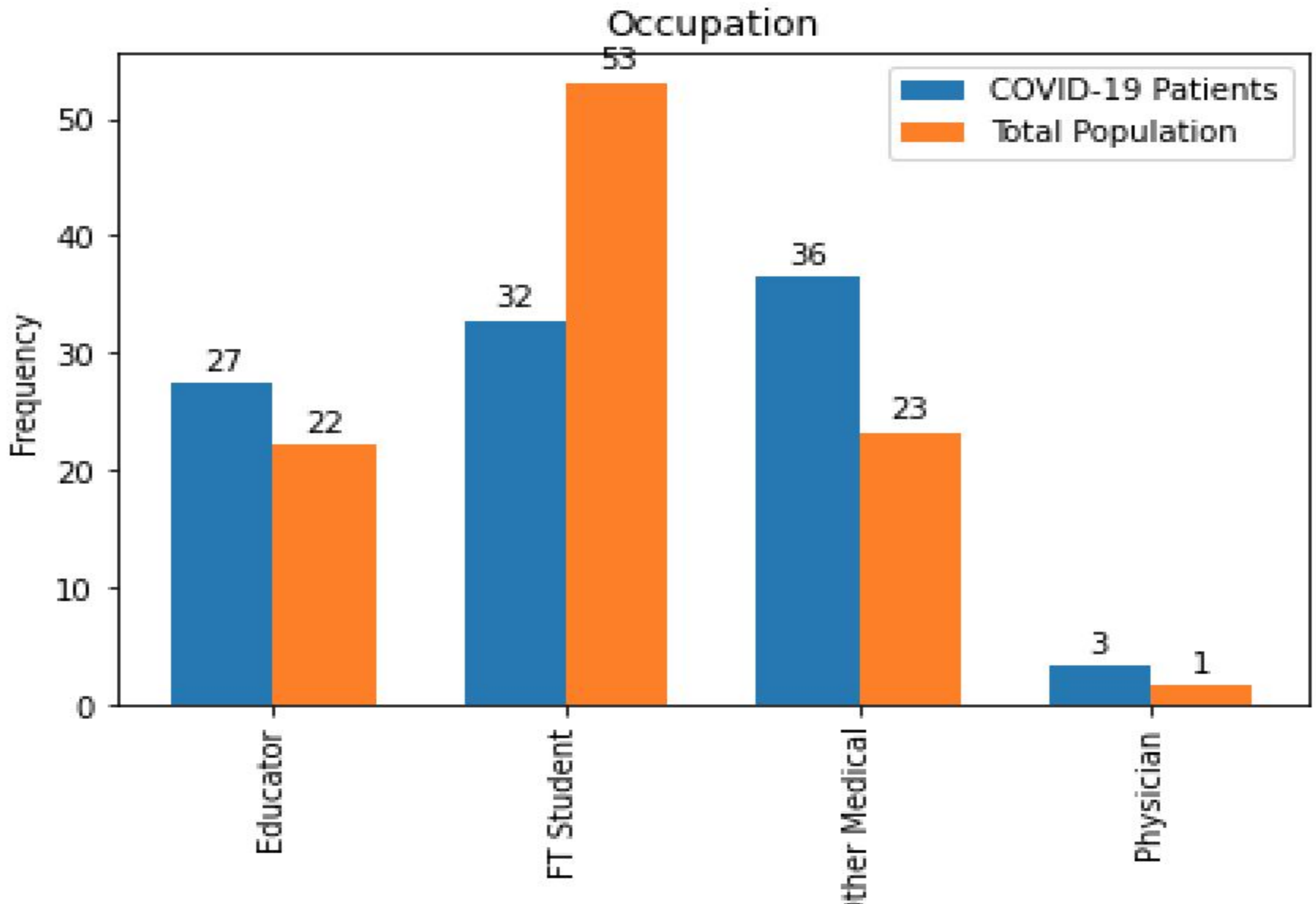
- Impact of different demographics
- Behaviors as risk factors
- Predictors (likelihood) as risk factors

Occupation*

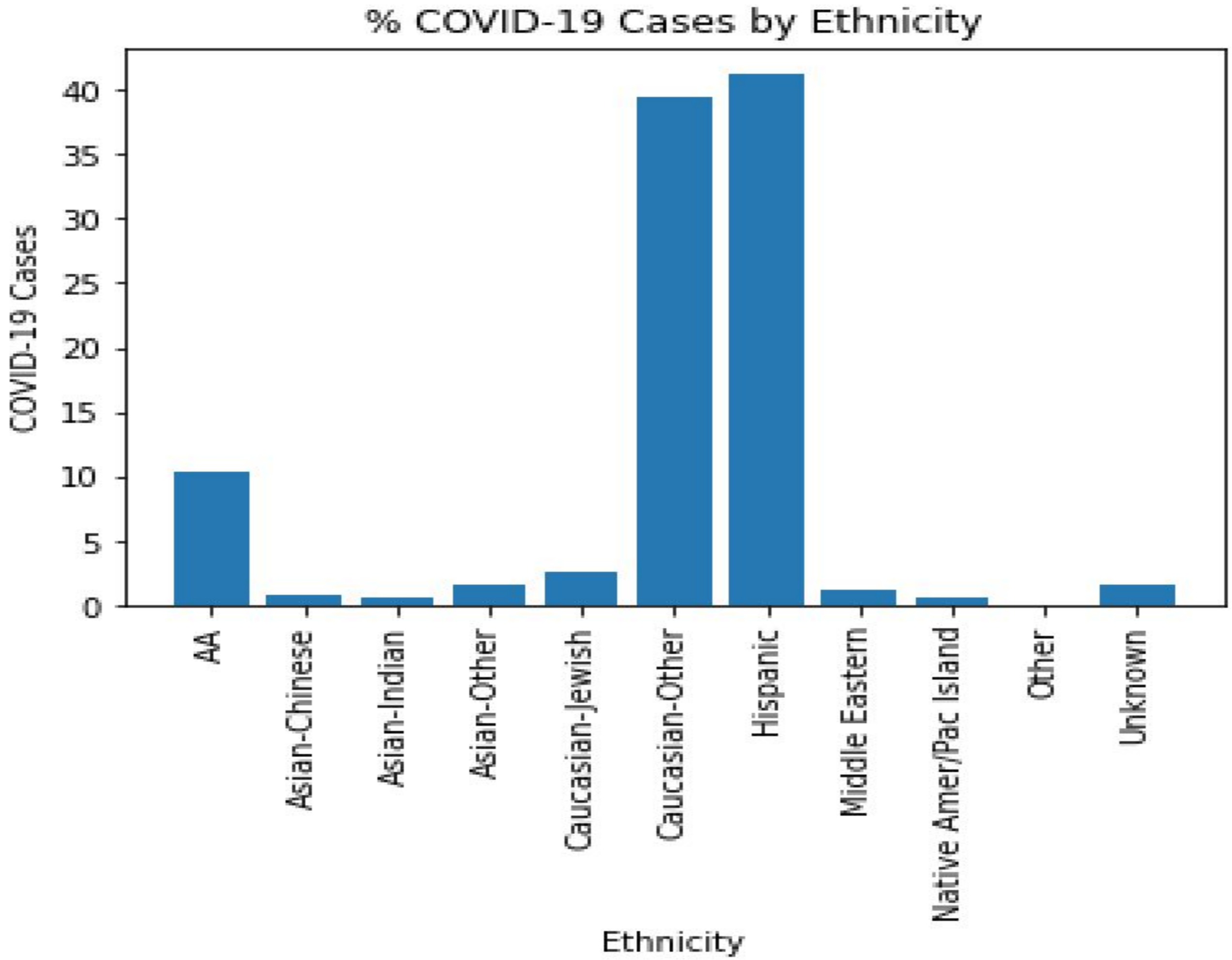


A3.Ai

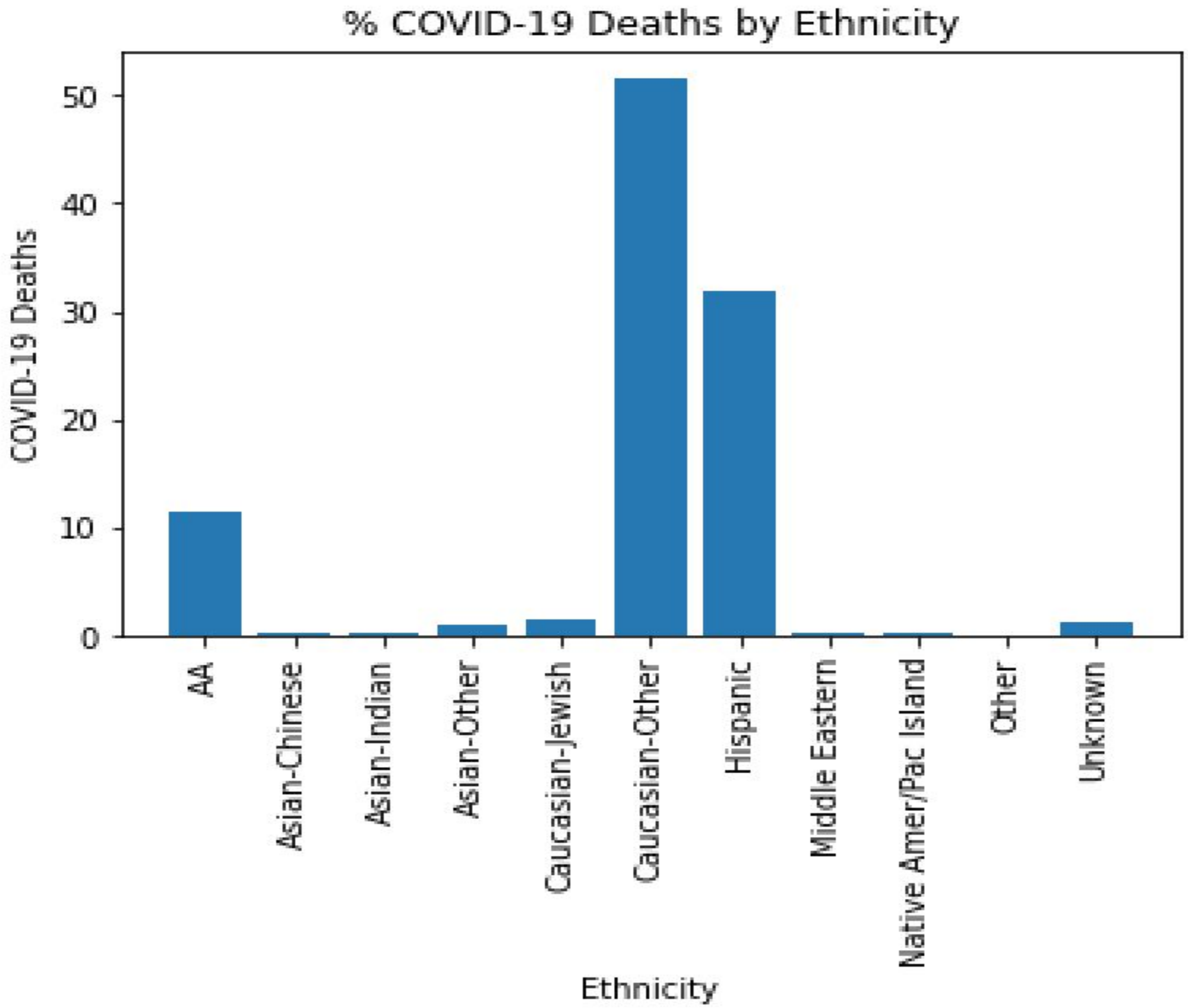
Specific Occupations



Ethnicity

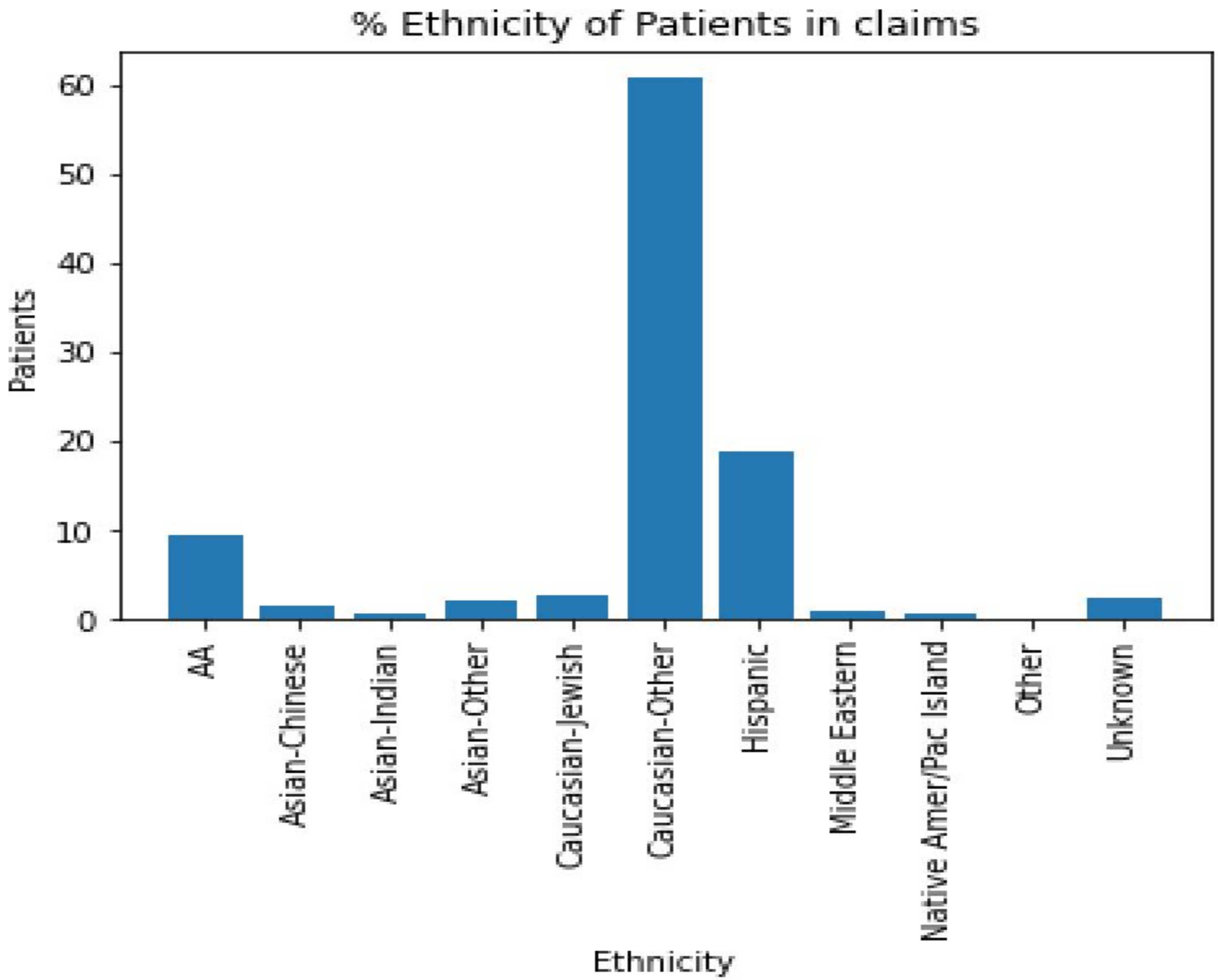

% COVID-19 Cases by Ethnicity
COVID-19 Cases
40
35
30
25
20
15
10
5
0
AA
Asian-Chinese
Asian-Indian
Asian-Other
Caucasian-Jewish
Caucasian-Other
Hispanic
Middle Eastern
Native Amer/Pac Island
Other
Unknown
Ethnicity

Ethnicity



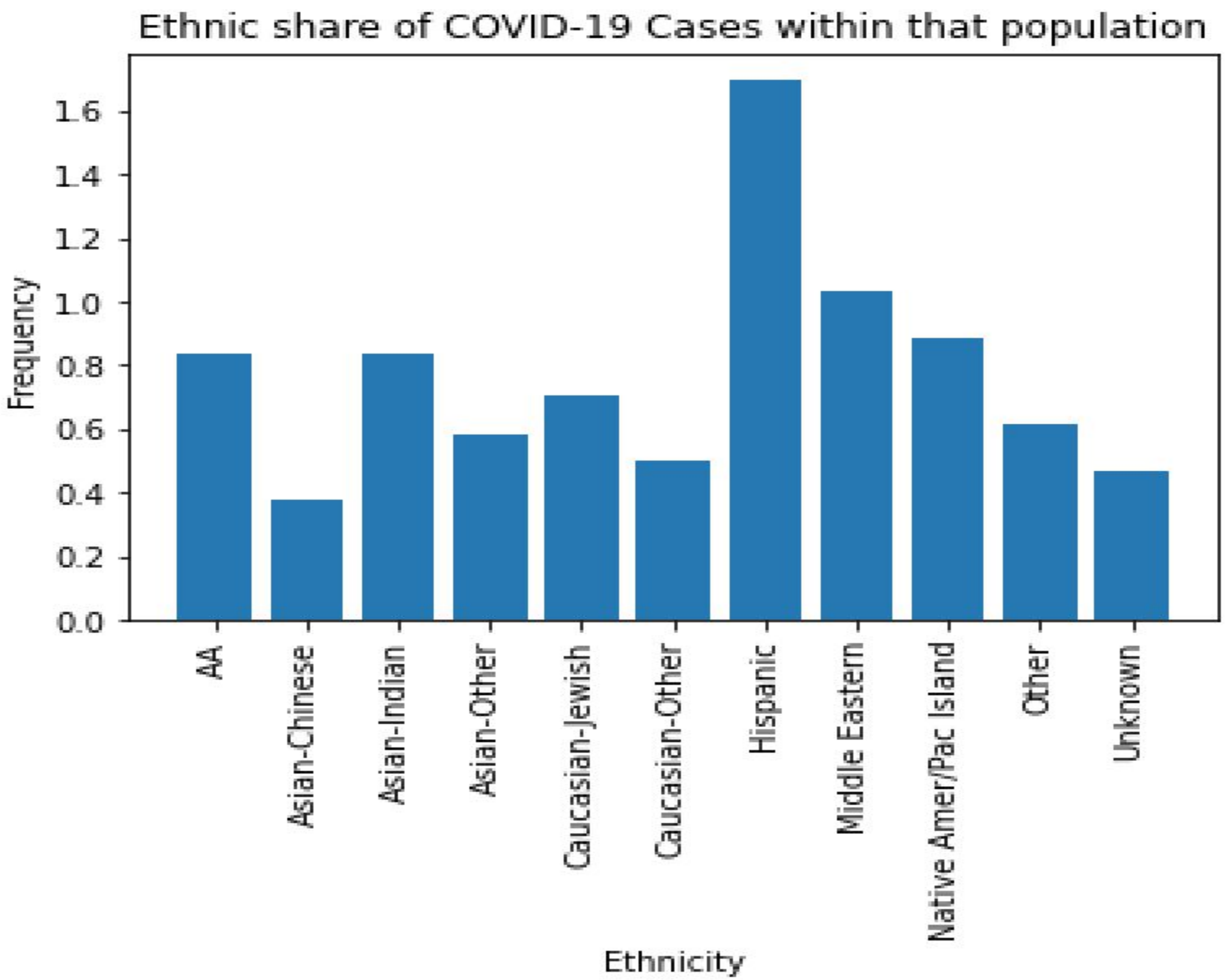
% COVID-19 Deaths by Ethnicity
COVID-19 Deaths
0
10
20
30
40
50
AA
Asian-Chinese
Asian-Indian
Asian-Other
Caucasian-Jewish
Caucasian-Other
Hispanic
Middle Eastern
Native Amer/Pac Island
Other
Unknown
Ethnicity

Ethnicity



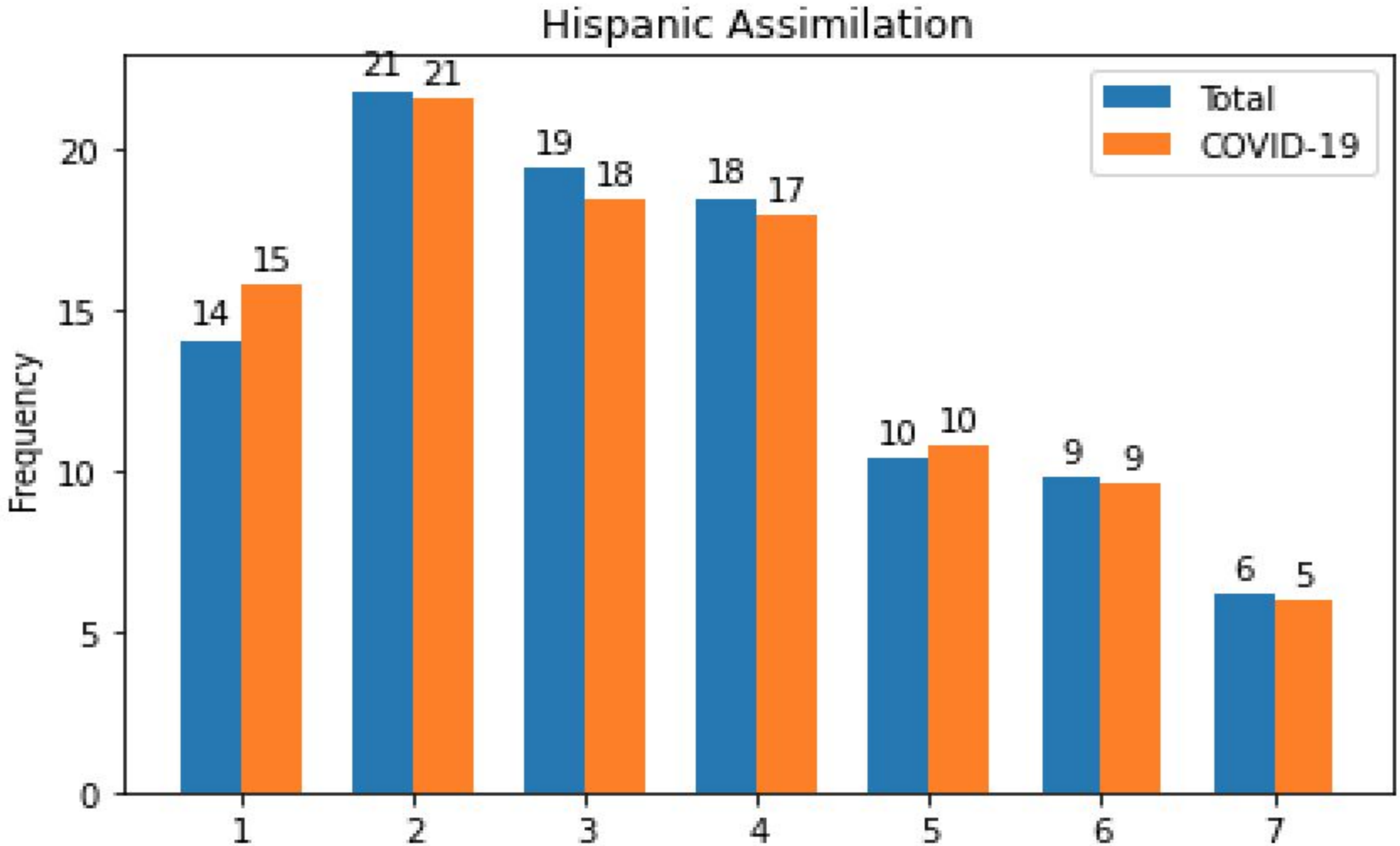
A3.Ai

Ethnicity



A3.Ai

Assimilation into US Culture



Least likely

Most likely

Future Work: SDOH

Many additional attributes available in the dataset:

BMI, Diet, Location, Profession, Access to Healthcare, behavior, etc

- Population Health study
- Predictive Models
- Personalized risk scoring

A3.Ai

Future Work: Current Use Case

- Improve model performance
  - Add clinical concept embeddings
  - Add temporal features (time sequence of clinical events)
  - Other machine learning and (lighter weight) deep learning models
- Add more classification categories (other severe outcome types)
- Incorporate Electronic Health Record data for prediction

Future Work: Population Health Dashboard

Adjustable thresholds cost, date range, etc

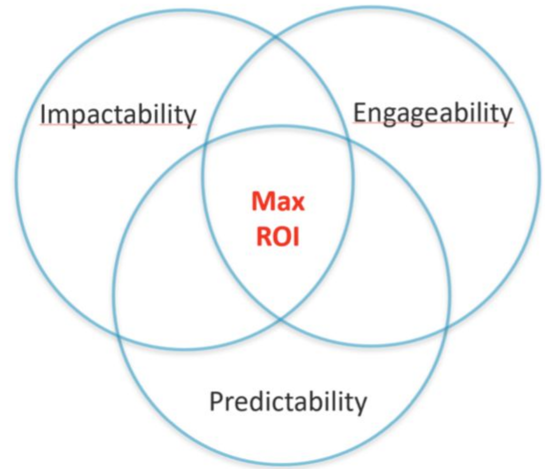
Filters by claims type, group account, etc

Key influencers/predictors: conditions, procedures, etc

#	Comorbidity	Prevalence
1	Cancer	0.18
2	Diabetes Mellitus	0.16
3	Chronic Pulmonary Disease	0.14
4	Chronic Kidney Disease	0.13
5	Congestive Heart Failure	0.11
6	Cerebrovascular Disease	0.09
7	Peripheral Vascular Disease	0.08
8	Diabetes Mellitus with complications	0.06
9	Metastatic solid tumor	0.05
10	Myocardial Infarction	0.05

Cost Breakdown

Population clusters by similarity



Future Work: New Use Cases As the Pandemic Progresses

- Drug and COVID-19 vaccine effectiveness and safety
- Identify clinical trial candidates
- Long term consequences of COVID-19
    - Personal
    - Societal
    - Health
    - Economical

Private AI

Collaborative Learning with Obfuscation Aggregation & Knowledge Transfer

Changrong Ji
David Patton, PhD
Vance Degen
Ben Carroll
Greg Ewing, JD

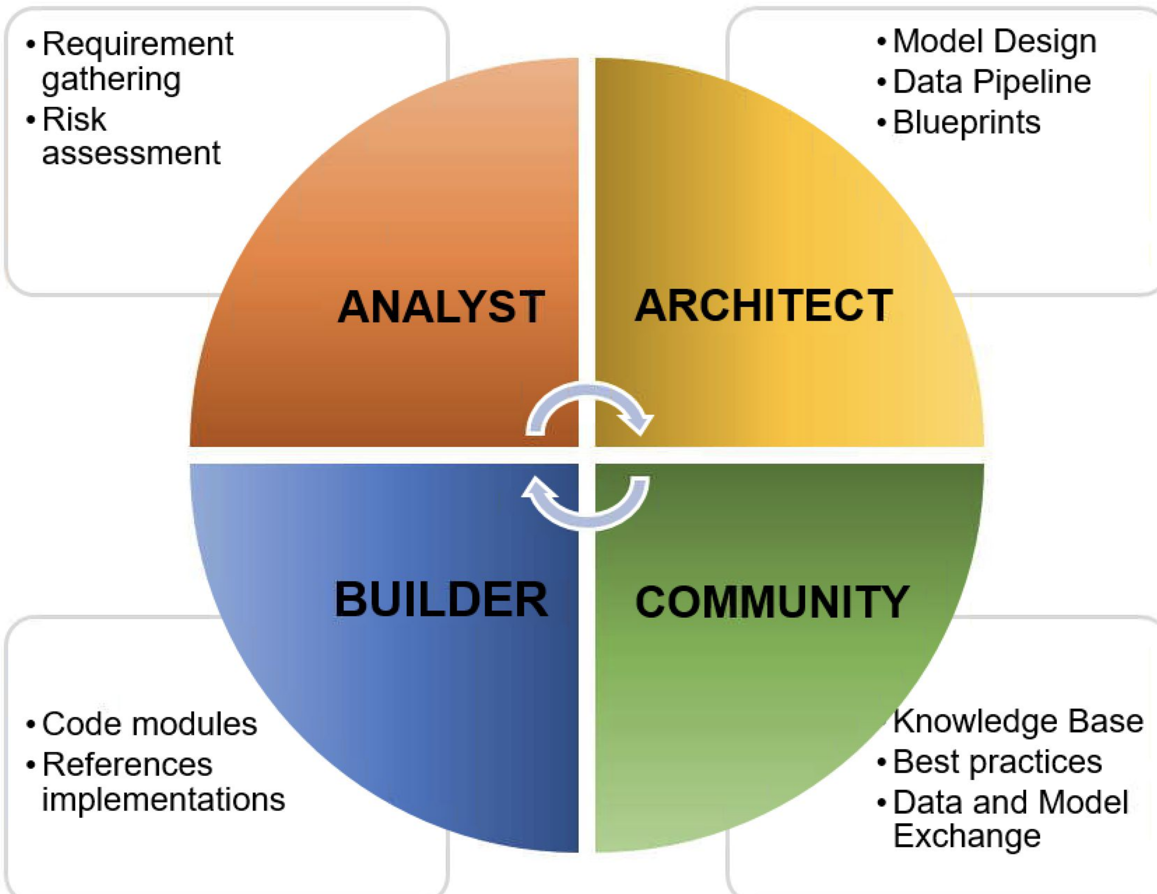
A3.AI

Data Sharing & Healthcare AI Challenges

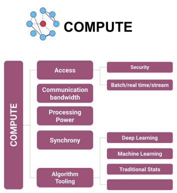
- Advanced analytics relies on data, often curated from multiple sources.
- Data sharing and usage have a complex web of trust with multiple parties: data owners, analytics solutions providers and users.
- Aggregating data and building models centrally raise concerns in
  - privacy and security,
  - single point of failure,
  - intellectual property and
  - misaligned incentives on the usage and value of of the combined data and models
- CLOAK is a platform and toolkit that is under development to address some key challenges in collaborative privacy preserving machine learning.

Specific relevance to the COVID-19 Research DB projects: The personal level medical records and social data, while de-identified, are still vulnerable to attacks such as data linkage and model inversion that leaks private information. The following highlights a set of techniques to mitigate the privacy risks. A series of papers will be published on this topic. Starting with: Privacy-Preserving Machine Learning Techniques 2020, Changrong Ji et al.

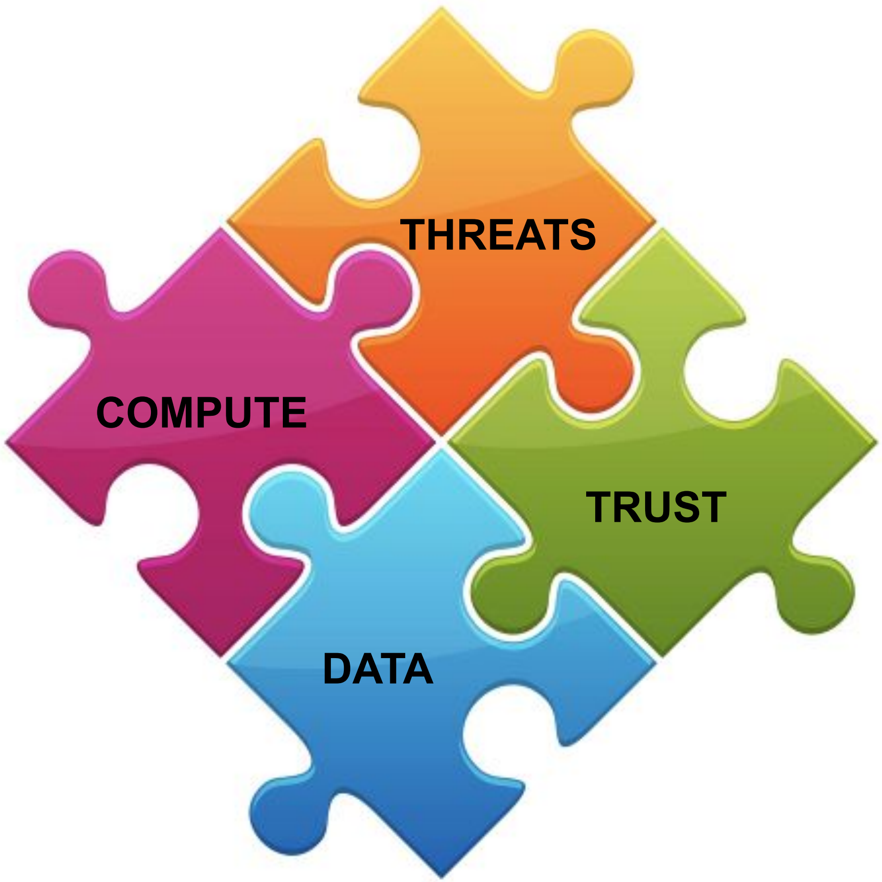
CLOAK PLATFORM (work in progress)



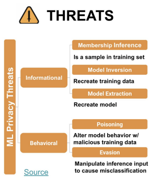
ANALYST



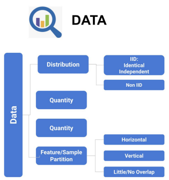
COMPUTE
COMPUTE
Access
Communication bandwidth
Processing Power
Synchrony
Algorithm Tooling
Security
Batch/real time/stream
Deep Learning
Machine Learning
Traditional Stats



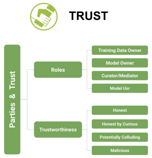
THREATS
COMPUTE
TRUST
DATA



THREATS
ML Privacy Threats
Informational
Behavioral
Membership Inference
Is a sample in training set
Model Inversion
Recreate training data
Model Extraction
Recreate model
Poisoning
Alter model behavior w/ malicious training data
Evasion
Manipulate inference input to cause misclassification
Source

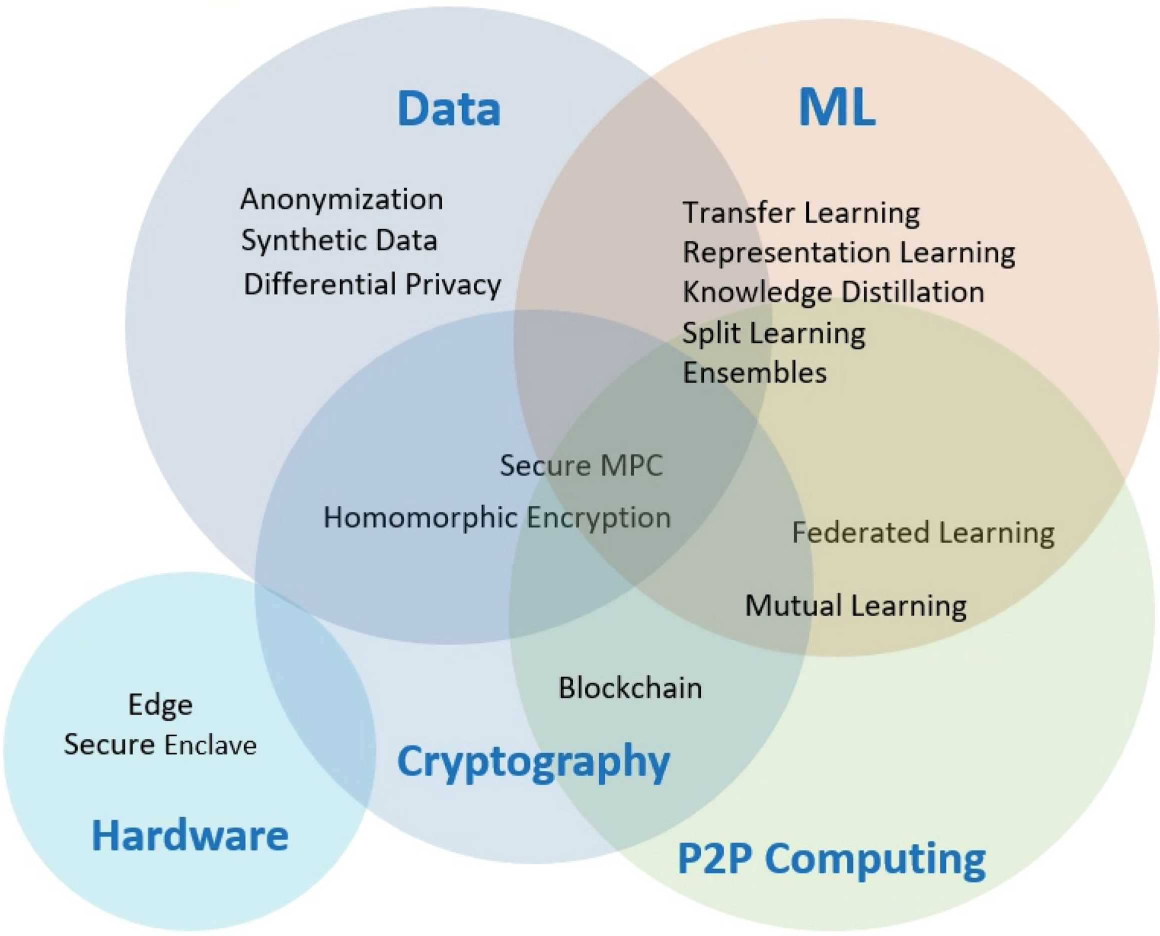


DATA
Data
Distribution
Quantity
Quantity
Feature/Sample Partition
IID: Identical Independent
Non IID
Horizontal
Vertical
Little/No Overlap



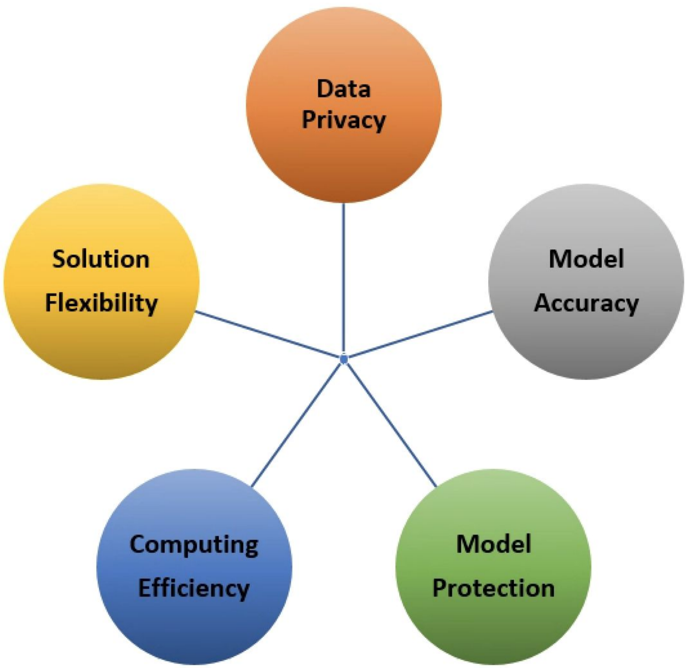
TRUST
Parties & Trust
Roles
Trustworthiness
Training Data Owner
Model Owner
Curator/Mediator
Model Usr
Honest
Honest by Curious
Potentially Colluding
Malicious

PRIVACY PRESERVING TECHNIQUES 2020

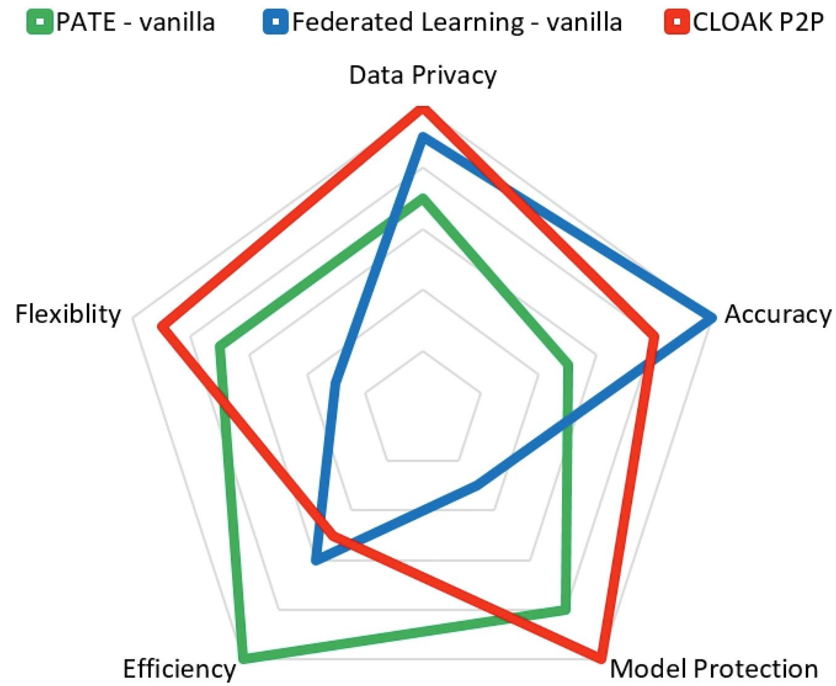


ARCHITECT

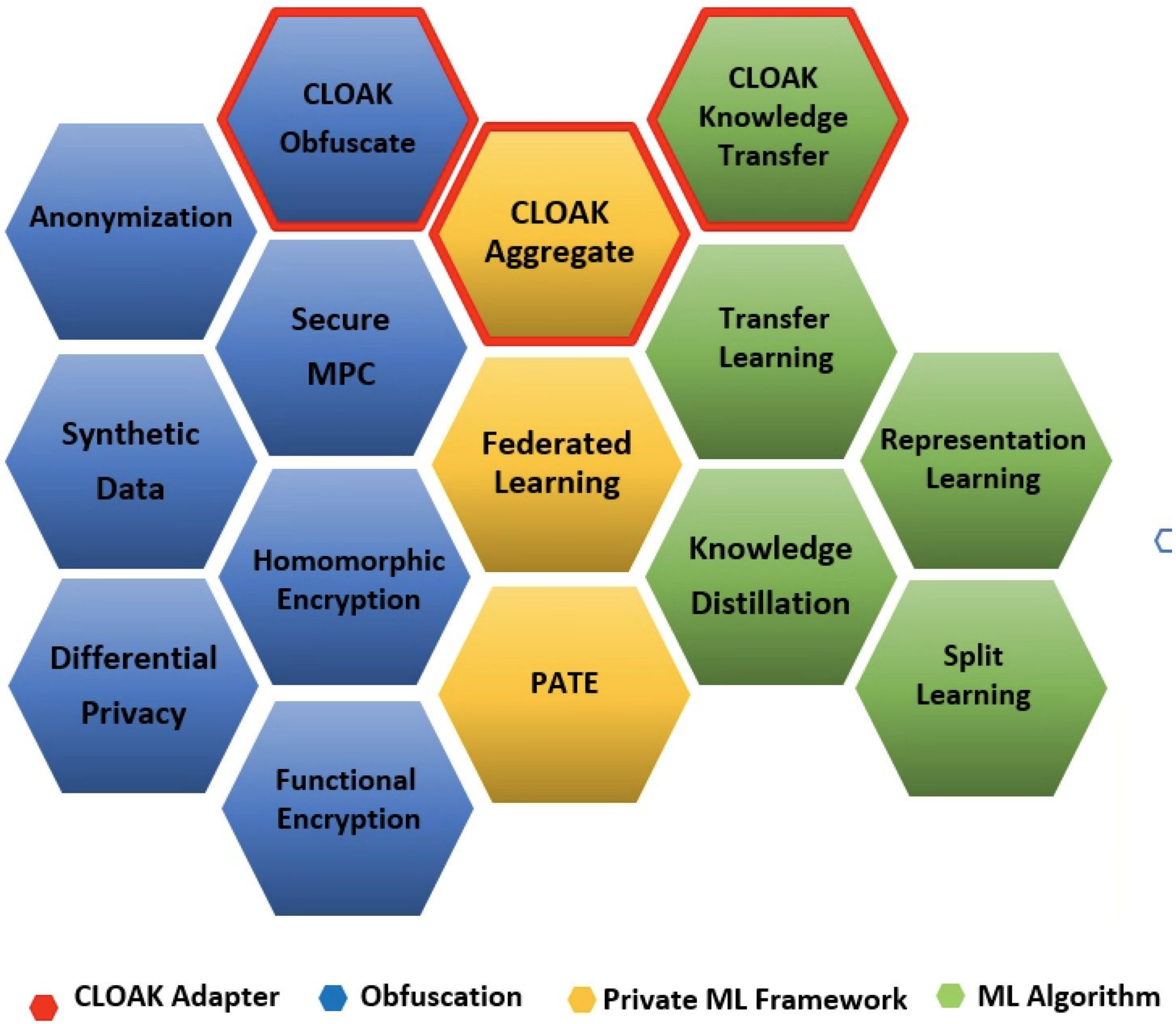
DESIGN TRADE-OFFS



EXAMPLES



BUILDER


CLOAK Obfuscate
CLOAK Knowledge Transfer
Anonymization
CLOAK Aggregate
Secure MPC
Transfer Learning
Synthetic Data
Federated Learning
Representation Learning
Homomorphic Encryption
Knowledge Distillation
Differential Privacy
PATE
Split Learning
Functional Encryption
CLOAK Adapter
Obfuscation
Private ML Framework
ML Algorithm