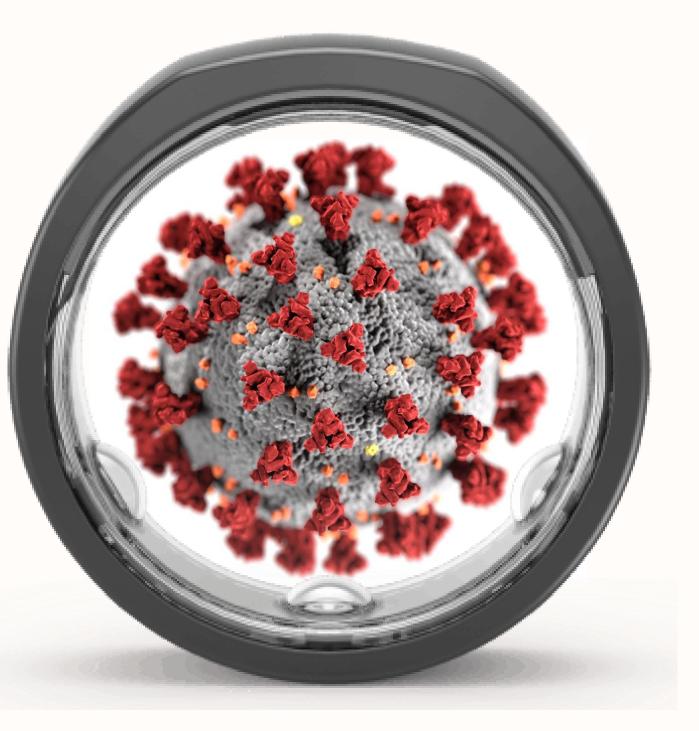
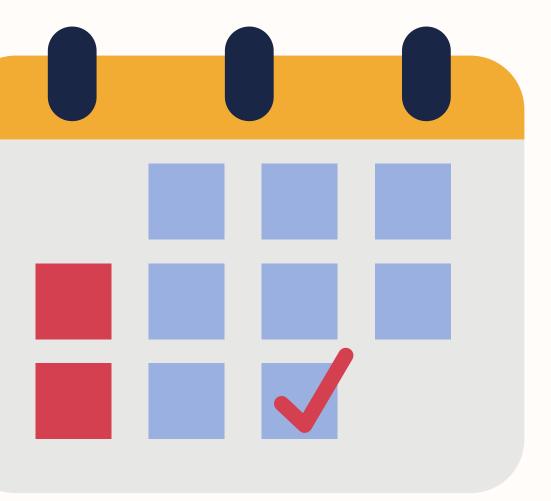
IoT Wearable Data-Fever Analysis & COVID Onset Detection

Wearables - Vital role to predict vital Information



- Team
- Project Overview
- Solution Architecture
- EDA (Exploratory Data Analysis)
- Data Preparation
- Modeling & Evaluation
- Scalability
- Visualization
- Video Demo
- Key Findings
- Future Use







### Swetha Varadharajan

Solution Architect



**Yogesh Bansal** Model Expert



ML Expert



### Team



### Sasi Mahalingam

#### Visualization Architect



#### **Prof. Benjamin Smarr** Advisor

## Project Overview



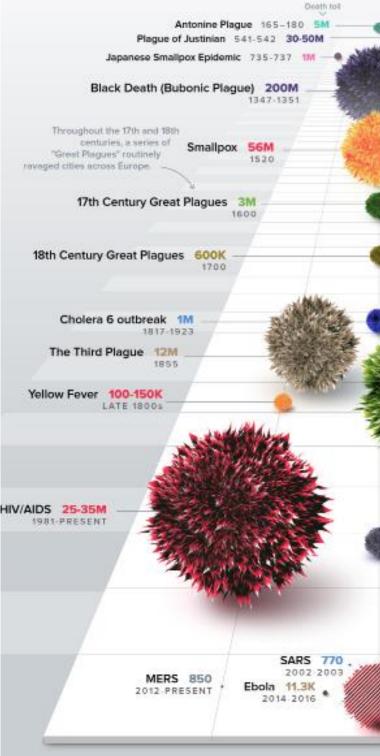


### HISTORY OF **PANDEMICS**

**Problem Statement** 

### Pandemics are not new to mankind

PAN-DEM-IC (of a disease) prevalent over a whole country or the world.

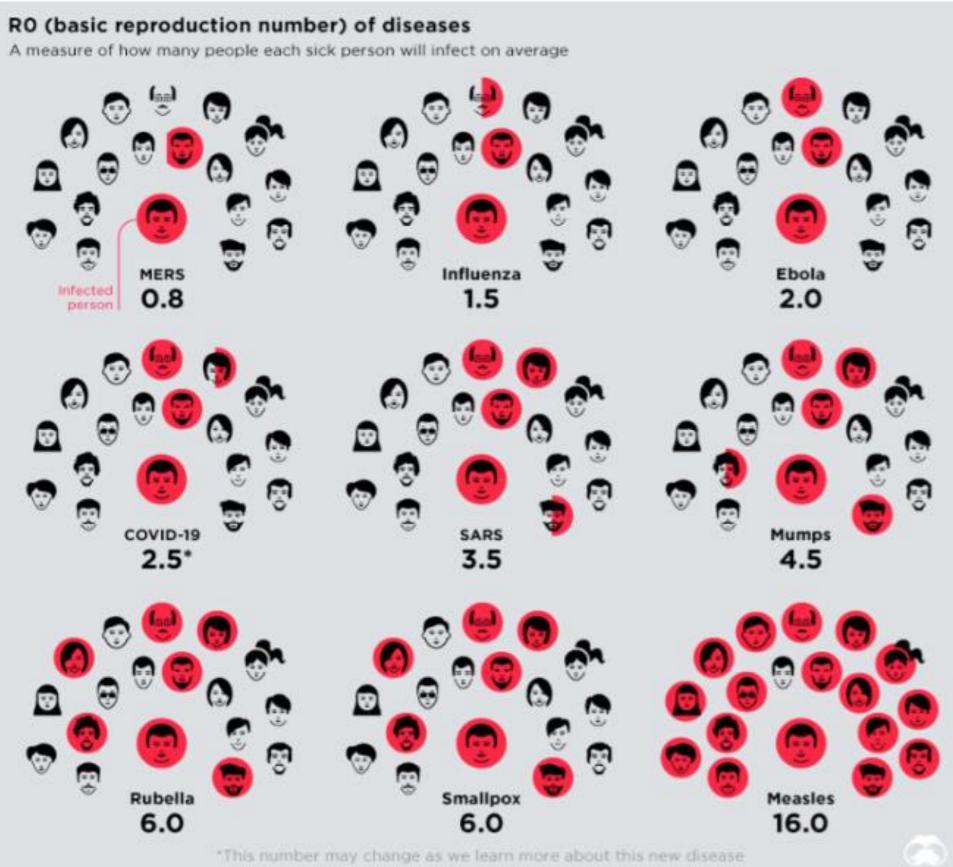


Source: https://www.visualcapitalist.com/

THROUGHOUT HISTORY, as humans spread across the world, infectious diseases have been a constant companion. Even in this modern era, outbreaks are nearly constant. Here are some of history's most deadly pandemics, from the Antonine Plague to COVID-19. 1450 1725 Spanish Flu 40-50M 1918-1919 Russian Flu 1M 1889-1890 Asian Flu 1.1M 1957 1958 Hong Kong Flu 1M 1968-1970 Swine Flu 200K COVID-19 3.5M\* 2019-9:20AM PT, MAY 27, 2021 [ONGOING] WHO officially declared COVID-19 a pandemic on Mar 11, 2020.

## **Problem Statement**

### Everyone has the responsibility to control the spread



Source: https://www.visualcapitalist.com/

### Corona Virus Status

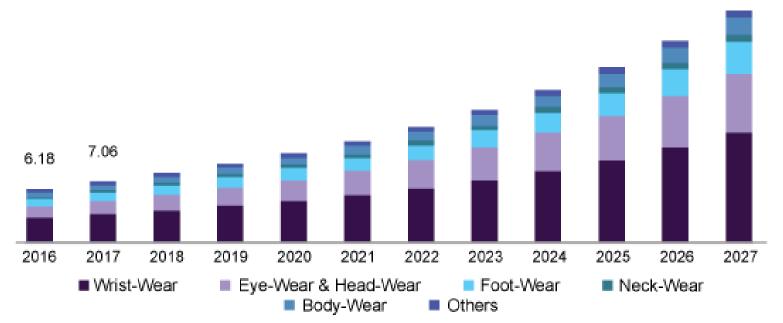


Source:https://coronavirus.jhu.edu/map.html

### **Data Resources**

#### Pervasive Wearables Industry

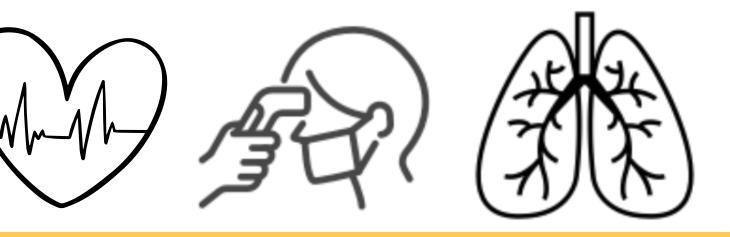
U.S. wearable technology market size, by product, 2016 - 2027 (USD Billion)



Source: www.grandviewresearch.com

### Physiological Signals from Oura Ring

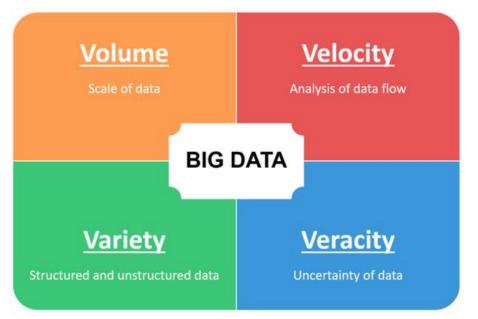




# Challenges

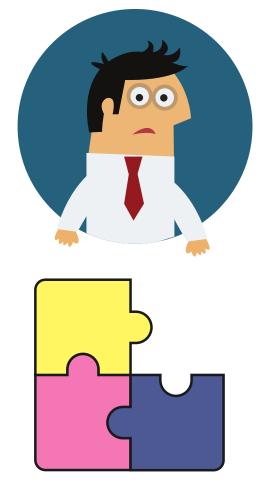


#### Missing Architecture

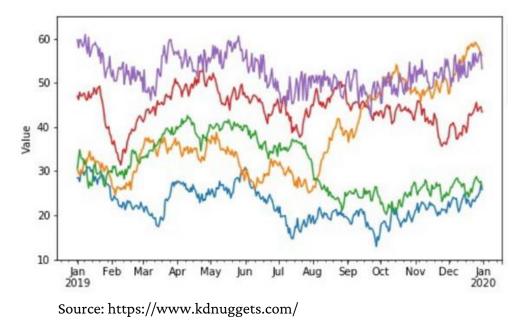


Source: https://chartio.com/

Big Data Management

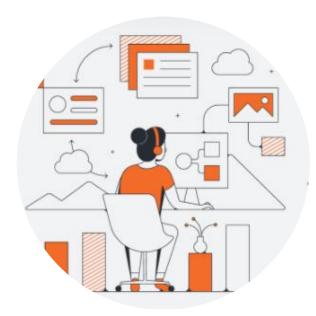


#### Missing Piece



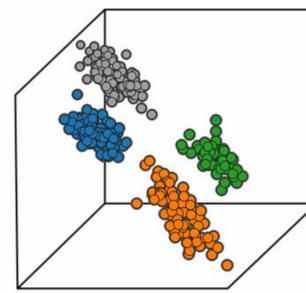
#### Time Series Data

# Opportunities



Architecture that enables rapid exploration at scale





Signal based Clustering



Covid Onset detection



Dynamic healthy baseline



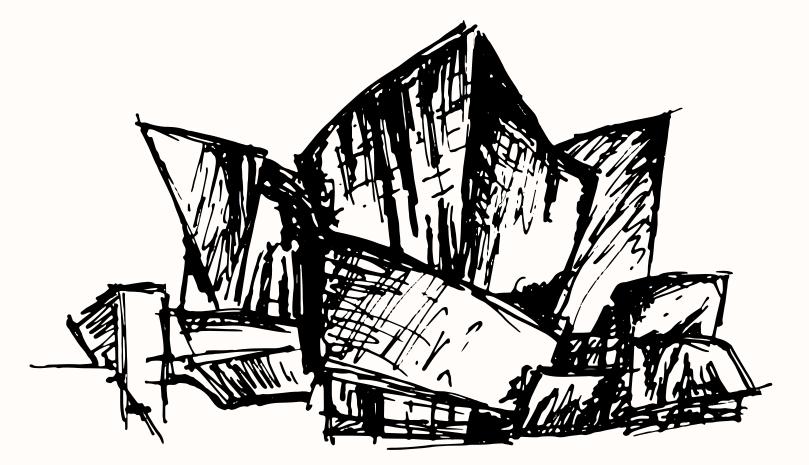


- Covid and Other Medical Researchers. i.e. Tempredict
- Wearable Health Tracker users.
- Data Teams working on harnessing physiological data.
- Domain Experts, Clinicians analyzing health sensor data.

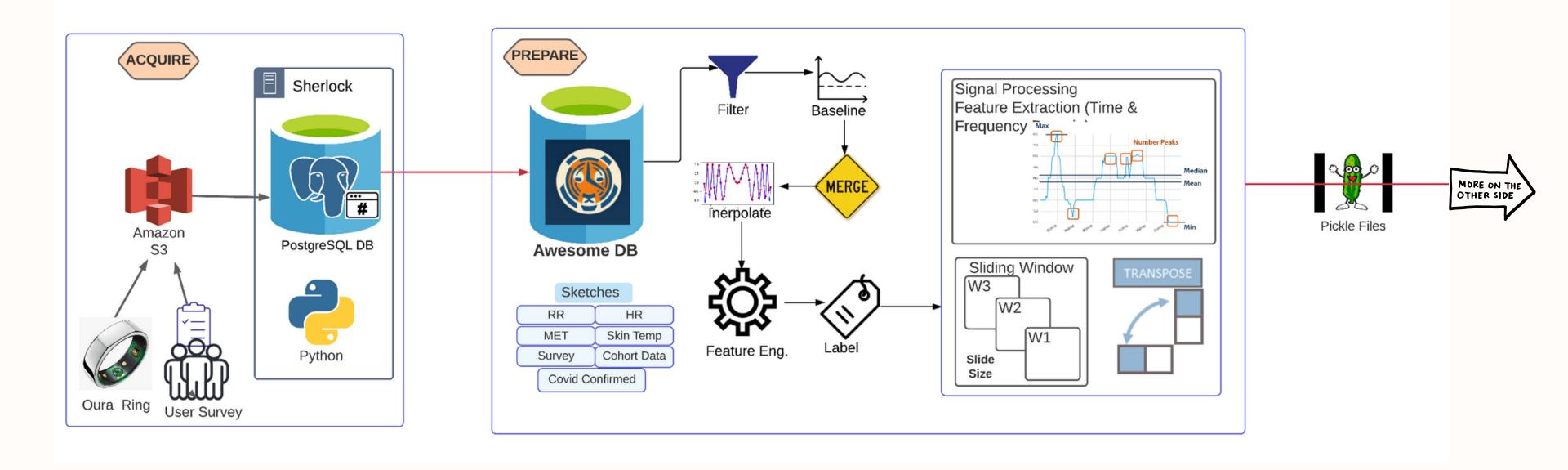




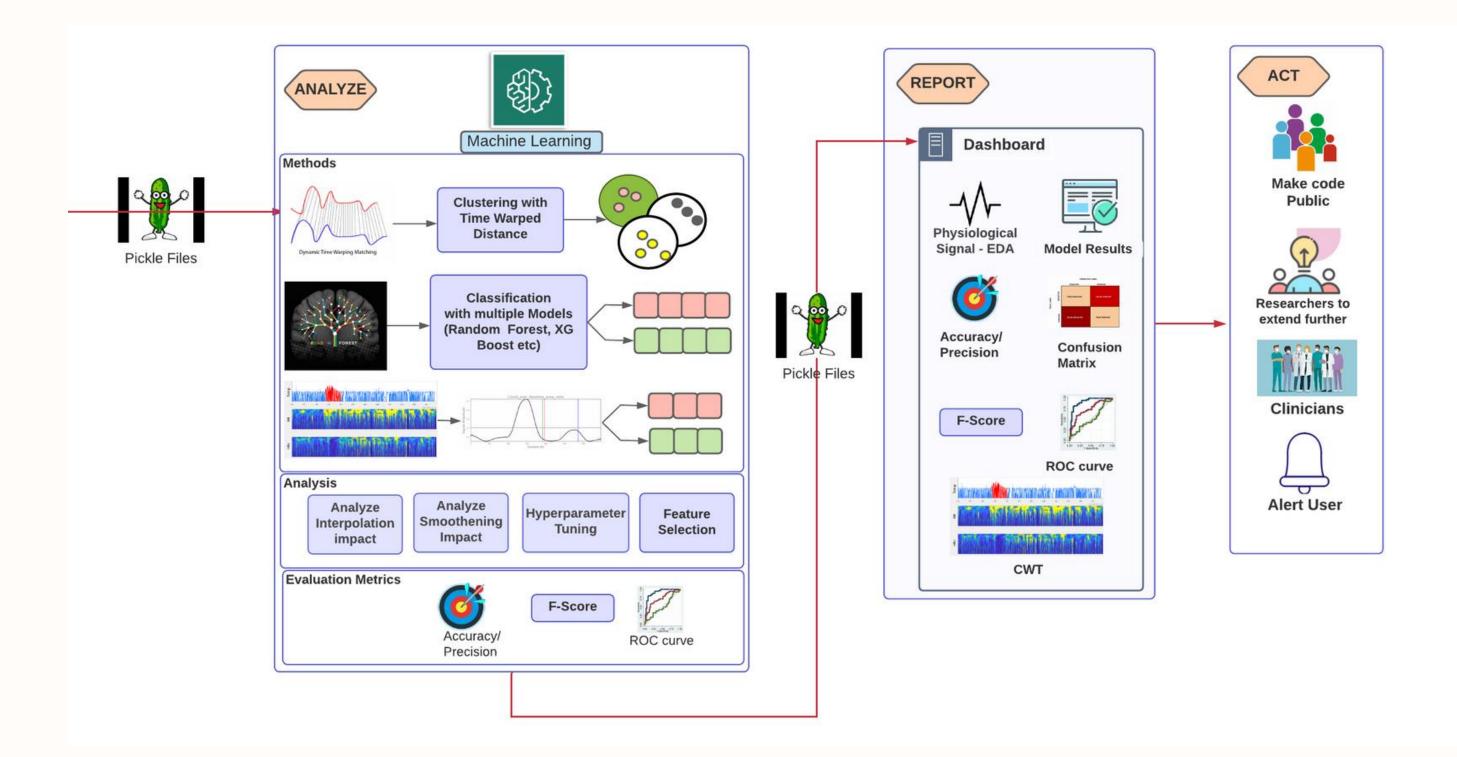
### **Solution Architecture**



# Solution Architecture



## Solution Architecture





### **Data Sources**

### Physiological Data from Oura ring

[Every minute]

- Skin Temperature
- Respiratory Rate (RR)
- Heart Rate HR (IBI -InterBeat Interval)
- Metabolic Equivalent for Task MET
- Sleep & Wake Pattern

#### Survey Data from Users

[Onboarding/Daily & Monthly]

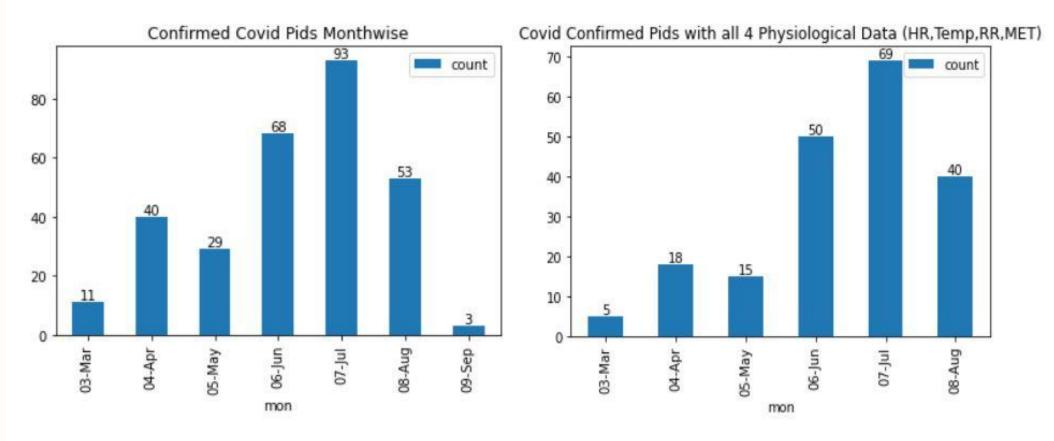
- Demography, comorbidities, Prescribed medications
- Symptoms [Dry Cough/Shortness of breath/Headache]
- Covid Test information (type and date)
- Other infections (flu, common cold, etc)

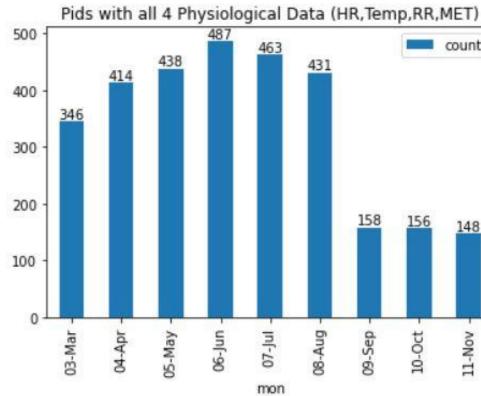


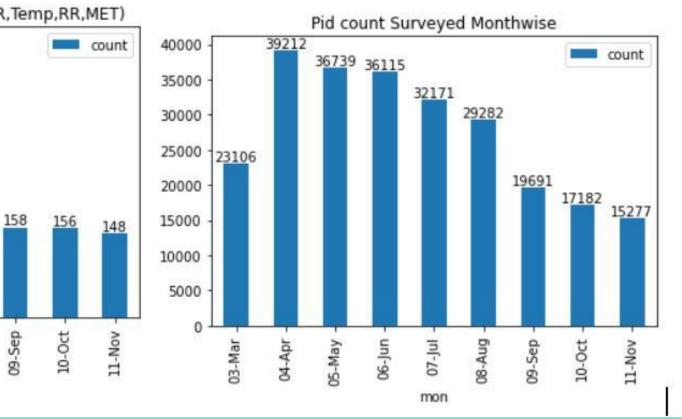


### **Data Statistics**

- Survey Data:
  3 Million surveys from 64K persons
- Device data from Oura ring: Minute by minute raw physiological data for 10 months
- PCR Confirmed COVID Cases:
  295 Individuals
- Golden Set 147: Based on data availability and signal rhythmicity





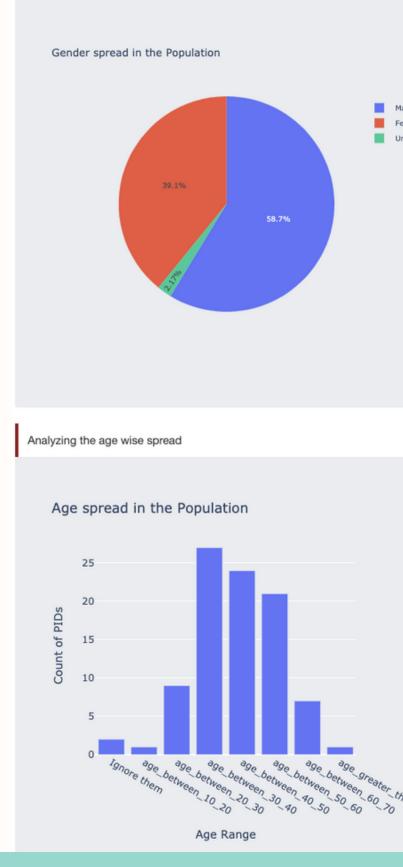


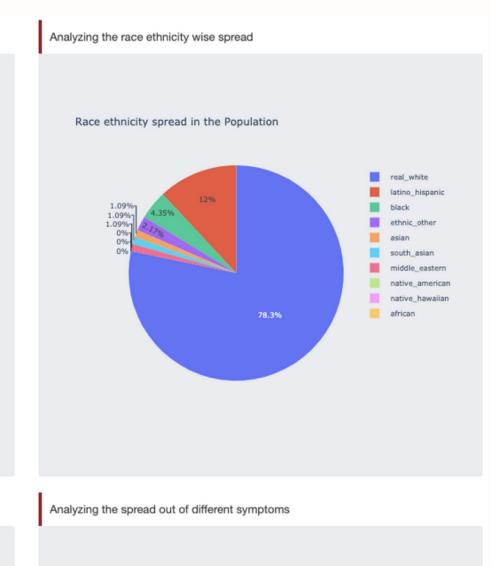
Analyzing the gender wise spread

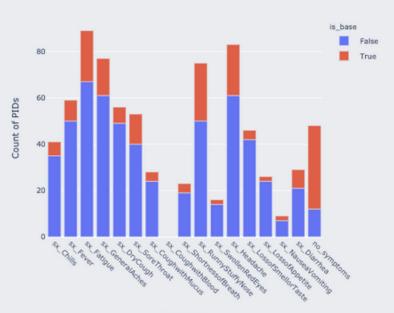
### **Data Statistics**

### Demography

- More Male Individuals than Female
- Majority Age range between 20 and 50
- Majority from the white population
- Presence of symptoms in both Covid and Baseline window



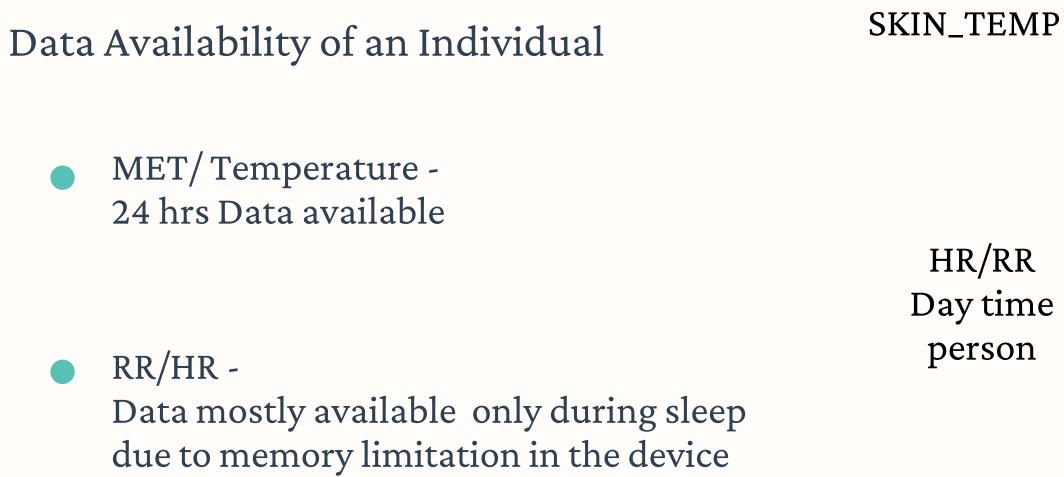


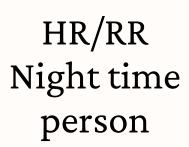


Variation in symptoms for baseline and COVID

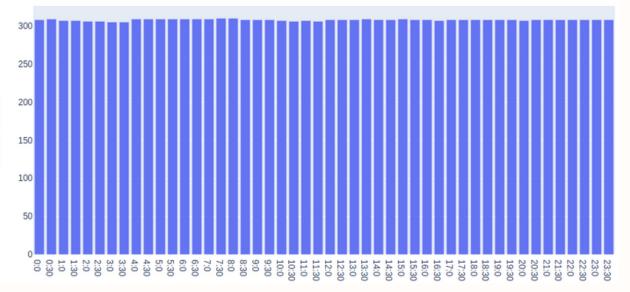
Male Female Unknown

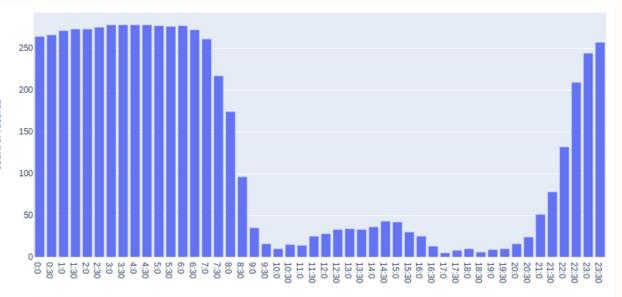
Symptoms

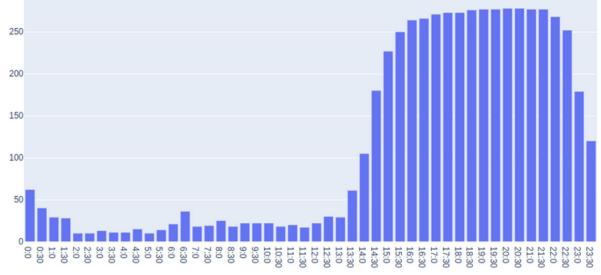




MET /



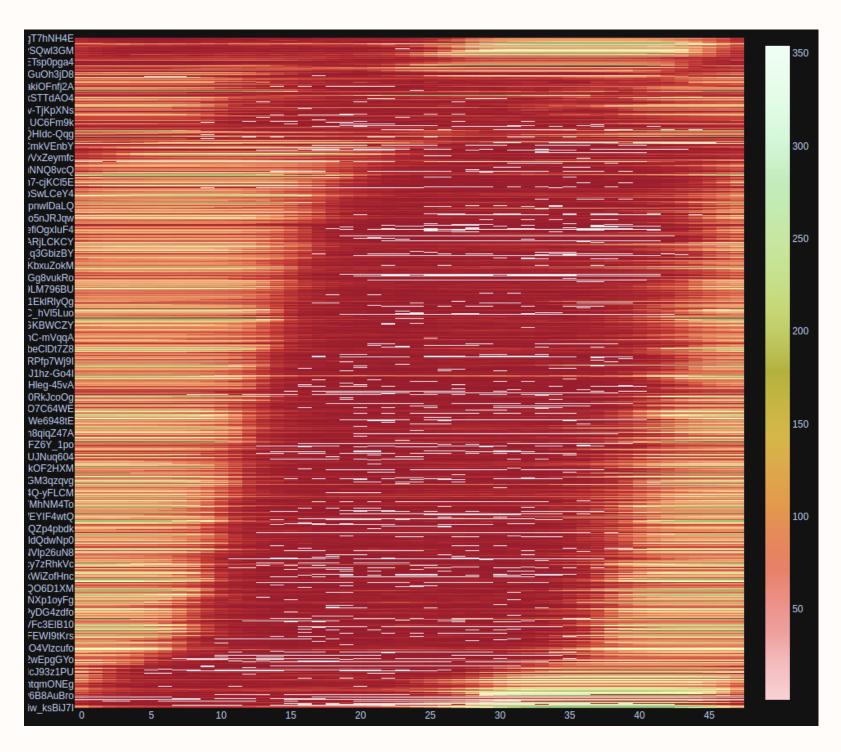




time\_of\_day

#### Data Availability of all individuals in the full dataset

over time of day					
thCe3epXmP5G47dr-UIEYIICq15xTN8YVQWsVSWdm1s					
KEWMKR0j4iF1vJpKYZD4rGaHq3rhg2PvswCod2sp2TGw					
OptokLNcf28Rn4KALhaM3818BpBJHnUyyR7KIMGcFAzg	-	-			
7ezU3YUszeYHxF0cw2SJAmg3KQKpmuzc-G8-WHYI-Qc	_				
W20HKVmbKYzwh_sHYTFTVSLdWCgi_jKKJMPLDLE-#Co					
4kjr2MnfSdGbG9nBcpTm5ZZFYTgaJMckLRePWmOjUlec					
I-C7urYegq5zZR07zHAK(WtHppyQV5ZkVQdpsUnOPWsY			_		
ROTsABD1F48ijLakoFEjST2Dfn3wmmk3ZlytvNXmn7q4874		_			
7h1/F2P3QmXK0dbKlA2etmAwtwlZz7u67ncLkDTrcNt5Gw					
vj3l4M2V3uh86yMo32aNqCjPme45a/VRKadUQ_IH/VFMg				-	
tzdnV2nu8Ovmpl3-B8781jal0abm7v2dVuGfpz-yOpP047Kc					
ZqcTJNAu_f_0ETu7jkcQ6LKD7iqqUzL_5a0C4A8enrlOxbM					
3pQ7KIPFHSymmkGVwdsUwOQoibpXKIkU3rlwloC48d5A				_	
cmyPvjnPmJ3SJYu6A9UfranPNIpJLartIpTErwUn8rU50IK0				1111 L	
ELDo-8hTxD6-hFzakVM3DuRohw9zUTNWA/UZU46 YXA					
QADRekDijM0ipcehbADCs263XH22bWmXG-s8h3Nyg					
ZHTC73JwkWLRiteJbeLSEvXgtX-IXgdEpmqP4sej5F770					
ENH7ztytKHB4st7268zc2ip-IRCwc2pu2VuCKOt2tlazGwpt0	_	_	_	_	
IPOkwFT9QKUPswchgOw0xKxKYBwVYxb3y928yEFS8d64					
rAJV7WGa4O67wiHuyLpJsJ4pCWcdBndLGtDecVh4mOg PfmrXPZWM1cyrQXG3a6ZZ_9dm2ggQ6BxQmhsK3n3Hk					
K-Yg91-WiaKbRi282J5ShebMmmhApmoZJmMBU_g8KA					
ilto852cm0VAXb92H3kPpOJ9CyKHYAosGV31DY0-be/MfQ \$JQ3BFOGUXGvGvfzEDm+HgyM3HaR1UKzNYEkKJn0f7g		_		_	
MEeedVpOwGl0e2RG43vbd_gXRocbWwLIKnKVxgeFgo					
Y/Mzxp-XG7X/s7GRehE0/6McR wWbLDR2EPcA61Y9M					
#SIQ5Z1kFqLTrR497FRv2RH0IPv7gpx8WwLPovsQYMs0					
OP-L2gEICUICdP1NYIU-gHtLWTtDJnh9uVBJi2h9NOgps					
FWNig3BVa2MpcdE6kG9RPAs1bSnsi8P0gWj9nE+SkinU					
W5xtp8NNhKFw8ICmUK0aTjs8tpCAMhL3yHMxxAyY2cEU					
XxXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX					
Zog/Aa-0CIFmb5L5yagPSR0g/2y1McISaH-u HQgU1Mbc					
HHoUWQDA/HEQLHÖQysAafRipXirpe6Fito3ivJjcX43VIA					
WSITL8Q49p8h93IMF5e2HjWsI-SW8ZReyjEK_g3Gb88Y					
8JzFaztjFqQcFwUA0J3qB8P0McxmLn5L9zV70C7dv7tsM			-		
AUpD2ybaCwqexgFtxSLyb6XWWE-9xDaPuTGtm1h7E8					
Pb4TE0xW33MkjU_d03LN-89hugszr43hi0Wcd8r4SHu9M					
GyyQecchm3NntX52kLLDr3IONk-VOdYPj8uhDPIUNv1vY					
AStdl2_wMURYgBGkLFWlpgrlzLPdbLEdyZxPylGBHdAec					
YHtrkNopnSCq99oC7ewsNjoCj3JNIAINQ1XKCe6qV0r-Y					
IHqdGDr2kR801eg0vh6E8Ep03B5bA7sxpwuON3QG6Mk					
AegdwbT4cZJKYa_WVQdTRmCVXx22_eDS_DtzlgAhd_I					
RGBntzo3GJG8F3UgnBM12x_xxq2FvLH8wYu8ISvvennQ					
fMMp1-eTDkF1IIzcv6QlCQbnokgy7NW6s6NbM5Q33Kn4					
S19C-wy4QU3buTrO1WXRS3sH2ssA08sy11gWmq9O02k					
2gTcTbyqf6Mqf8Uat_EF2tHm2t5S3OvegH9e_N15YWxTU zmUCQ3AwatBhPY2MILWFingtswqP3tC_WAji2OYuePN04					
whvrF08F4EDFgcjr5ungby_qu72P_GmDQljGCak570n_2t oynFZILbRxvhcpLSwkCxmhrdguOu7IkIE9RSP5ax1IC1YQ					
JHLnOrCkDXxxxxeBUgIMgqBuvQZ3iCubudJN-tHCZG2yci					
ByNNFBNZu_2n4C5tVpX568pgp41WA3WhOC4wNrKA					
R6NisS_5HR182Lk0immge4hcvqVbZBVcmhYev9nRs8U		-			
aDMBclBtN4n.h1CXpJMXLWTs9shq8seKy3BX83tS6R8o					
MYILM/RC220/Geltz-brOdWDX-oxEtt ITar/WEISB8SwEV/					
L3Rz9qh6gKJwLgNK49MwymsJiaBN5EsiztweLbIOGHcM		-	-		
0	10	20	30	40	



Baseline Window The time period when the person is healthy

#### Duration

3 weeks to account for Weekly and daily rhythmicity

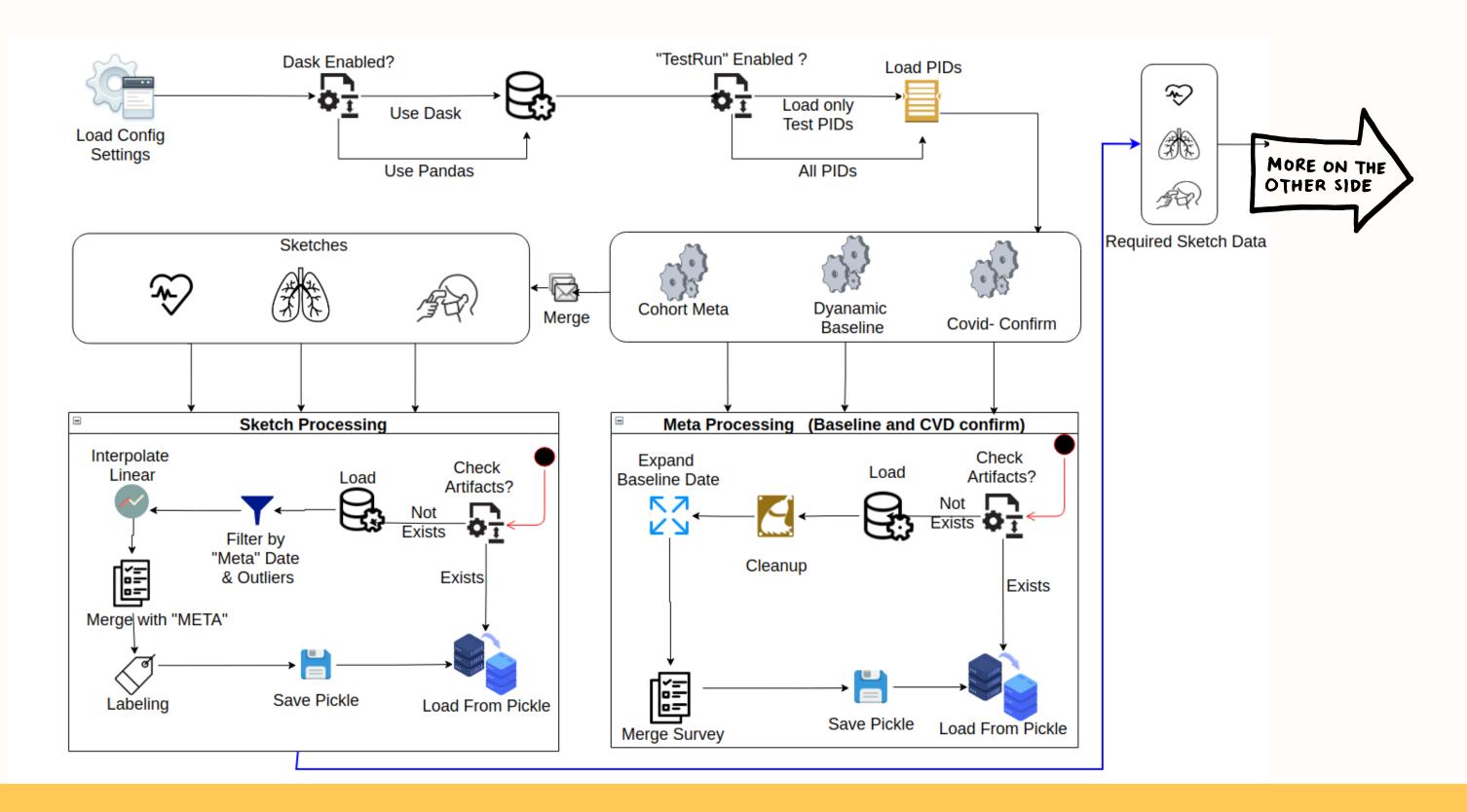
- Covid Window The time period when the person is infected or to detect infection
- Duration 3 weeks

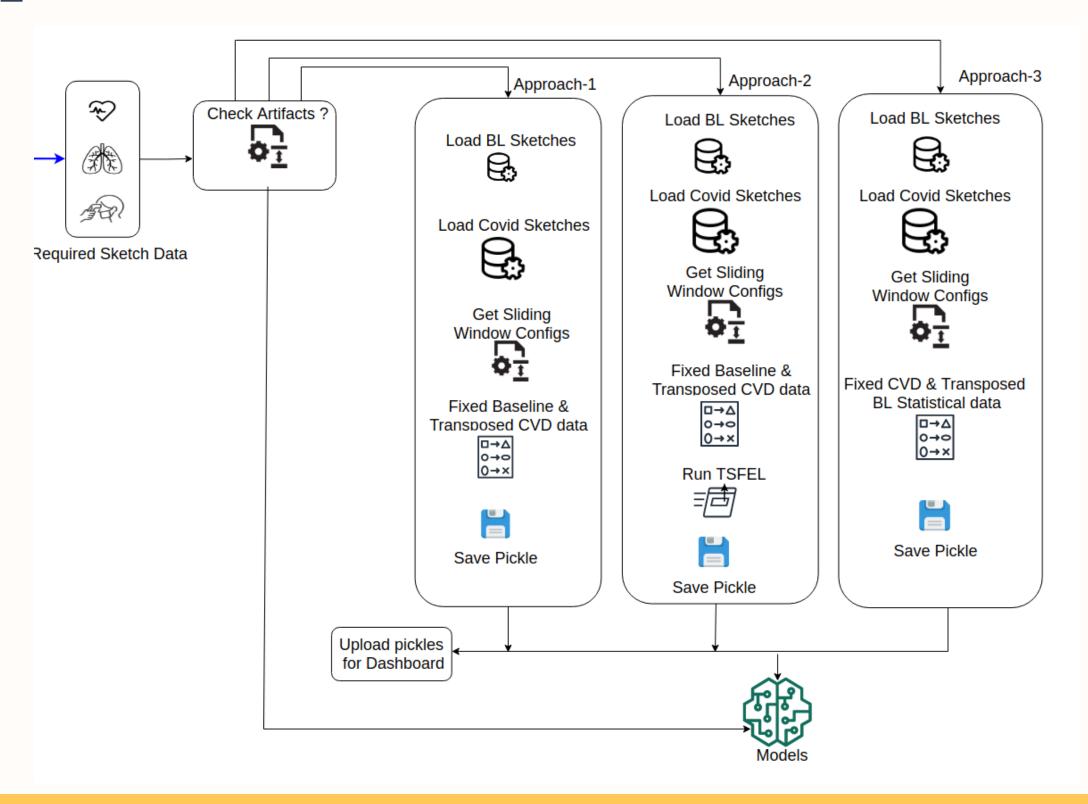












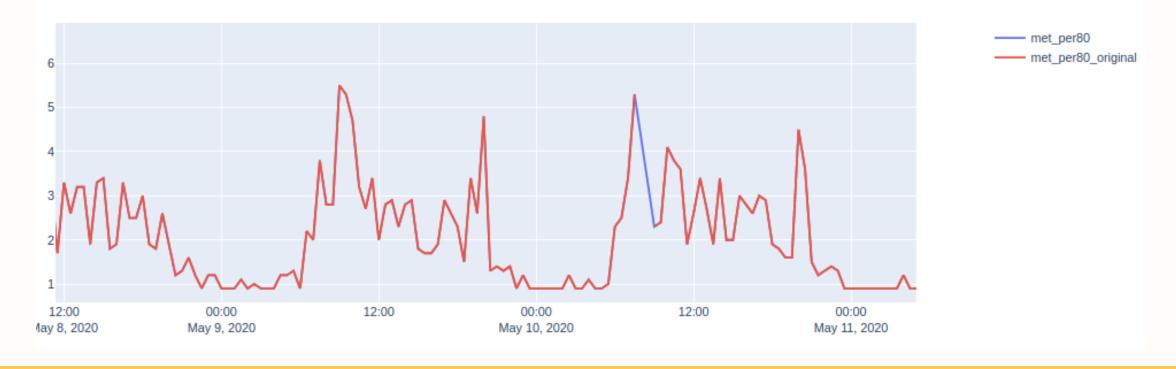
### Interpolation

Linear interpolation over time for the missing values preserves the rhythm of the physiology data.



• Filter Outliers

Remove Nonwear times Remove unrealistic recordings



#### MET Plot

Baseline - 21 Days (3 weeks)

- By Days of the week -3 Sundays, Mondays, etc..
- All Available continuous 7 days
- 30, 60, 90 days baselines

-68001-D:0-84DIO6-CD-VI	I-NTIDTI	Var Ome OV as COLLAR
gosQOIaDiOps4PJOoxOPa1J	lgNTgrJDTket22DS-eghguMJPLLj	- v xrQnir8 1 c038n40
hoose HRV feature		
× hr_per80		
1h 1d 1m 6m	YTD 1y all	
62		
n and		
47.5		
	· · ·	• •
Jul 12 2020	Jul 19	Jul 26
Manal	- Man	
a a construction		



# Clustering

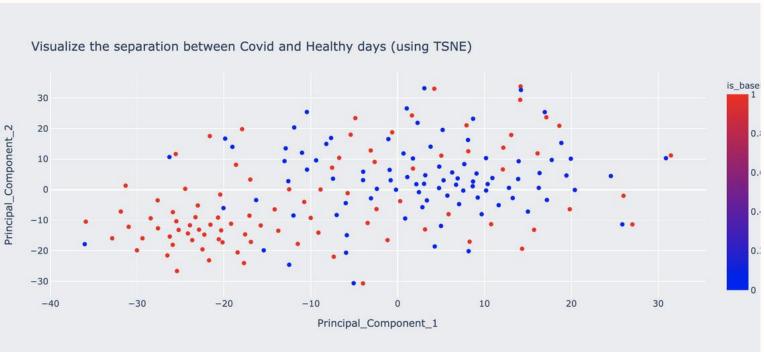
Cluster individuals based on variances in Temp, MET, HR, RR, and RMSSD. Transposing them into 3200 features

3200 Features were reduced to 450 using PCA

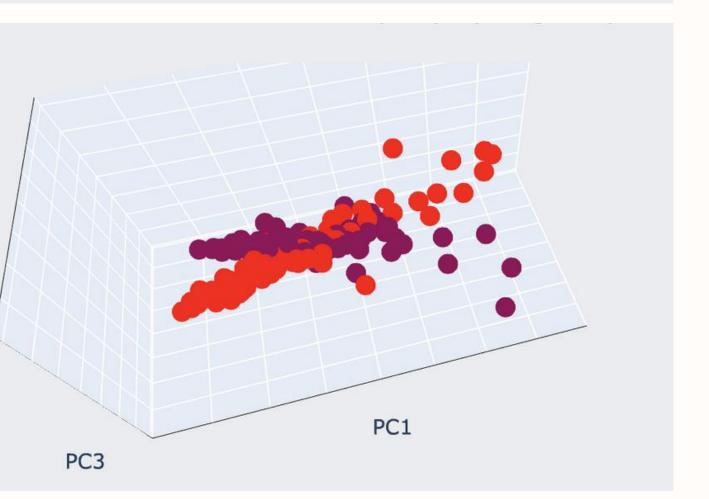
Gaussian distribution clustering with cluster size 4 to produce true covid and baseline clusters.

To visualize in 2D/3D the cluster input (450 dimensions) is further reduced to 2/3 using TSNE and PCA.

60% are clustered in different clusters during covid and baseline



PC2



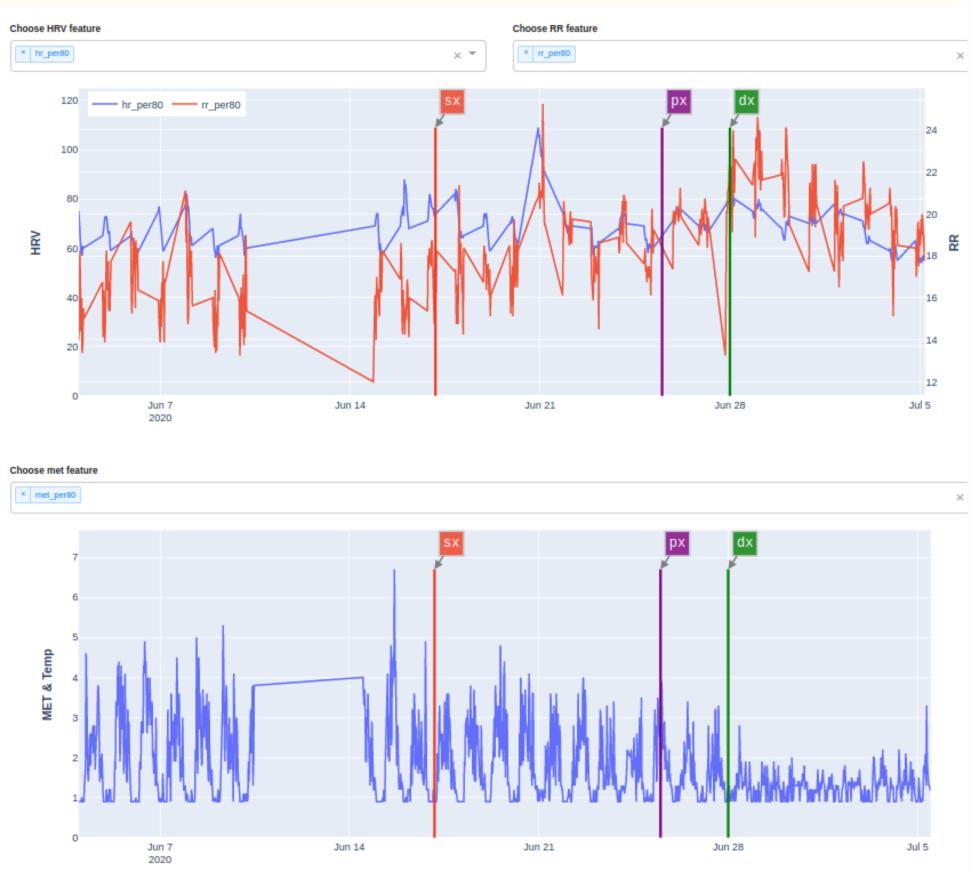
# Labeling

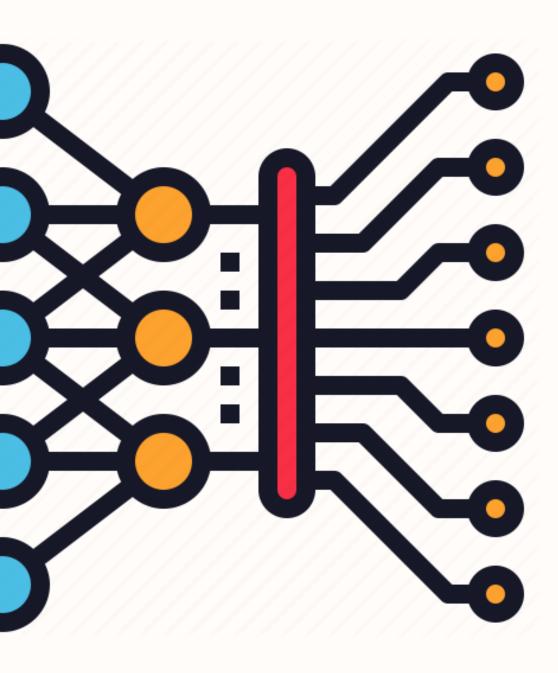
- Sx Date: Symptom Onset Date
- Dx Date: Diagnosis Date

• Px Date: Median Date between Max of HR Variation date and Max of RR Variation date

#### Target Labeling

+/- 2 days between Px date and Dx date (Diagnosis date)





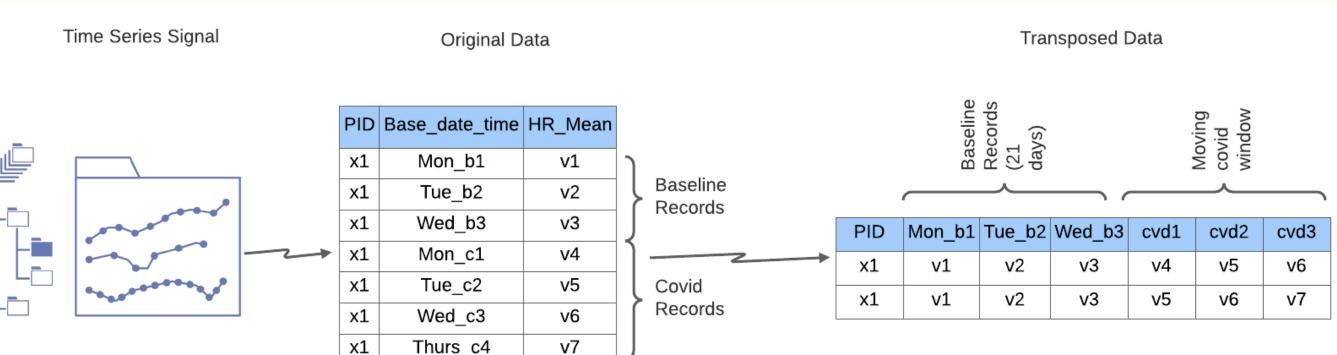
### **Approach-1:** Transposed Baseline & Sliding Covid Window

For each physiological data, 21 days baseline, a sliding Covid window of 3 days was taken.

Data was transposed to columns and passed as input and train the model

Target : Covid Onset (early detection)

Input data shape : 1530 \* 10518 Total Population:90 Train Test split : 80/20



#### Approach-1: Feature Importance & Results

0.0025

Random Forest Classifier Classification Technique Feature Importance

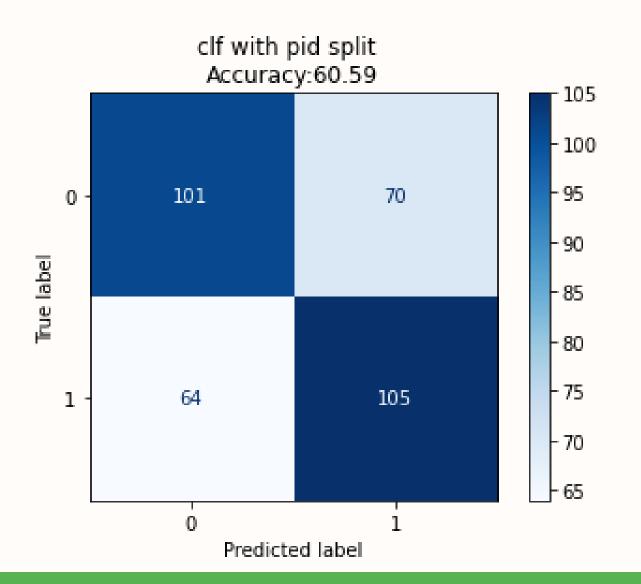
Visualizing Important Features skin\_temp\_per80\_cvd\_day\_2\_T\_02:30:00 skin\_temp\_per80\_cvd\_day\_1\_T\_00:00:00 skin\_temp\_per80\_cvd\_day\_1\_T\_01:30:00 skin\_temp\_per80\_cvd\_day\_2\_T\_01:30:00 skin\_temp\_per80\_cvd\_day\_1\_T\_23:30:00 skin\_temp\_per80\_cvd\_day\_2\_T\_02:00:00 skin\_temp\_per80\_cvd\_day\_1\_T\_04:30:00 skin\_temp\_per80\_cvd\_day\_2\_T\_00:30:00 skin\_temp\_per80\_cvd\_day\_2\_T\_03:30:00 ຍິ skin\_temp\_per80\_cvd\_day\_3\_T\_21:30:00 rr\_per80\_cvd\_day\_1\_T\_01:30:00 met\_stddev\_cvd\_day\_2\_T\_16:30:00 skin\_temp\_per80\_cvd\_day\_3\_T\_03:30:00 met\_stddev\_cvd\_day\_3\_T\_13:00:00 skin\_temp\_per80\_cvd\_day\_2\_T\_23:00:00 skin\_temp\_per80\_cvd\_day\_3\_T\_20:30:00 skin\_temp\_per80\_cvd\_day\_1\_T\_07:00:00 m\_per80\_cvd\_day\_1\_T\_03:00:00 rr\_stddev\_cvd\_day\_2\_T\_04:00:00 met\_variance\_cvd\_day\_3\_T\_14:00:00 0.0005 0.0010 0.0000 0.0015 0.0020 Feature Importance Score

#### Comparison between different algorithms

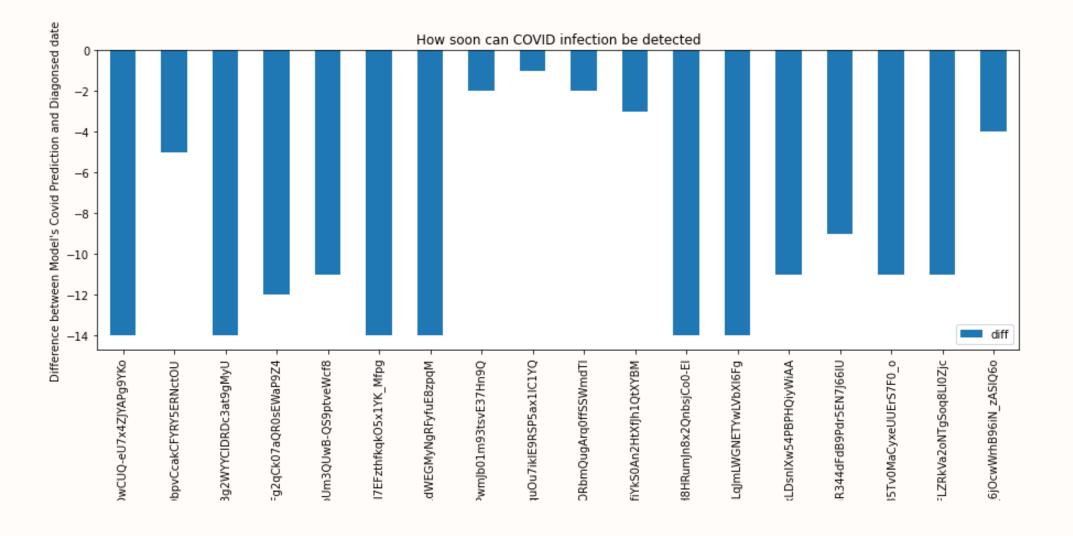
APPROACH-1					
Test train PID split	Model/Classifier	Accuracy	Precision	Recall	Fscore
(80:20)	RandomForest	57.84%	0.5893	0.5784	0.5752
(80:20)	XGB	55.23 %	0.5635	0.5522	0.5477
(80:20)	Bagging	56.86 %.	0.5854	0.5686	0.5611
(80:20)	GradientBoosting	53.27 %.	0.5465	0.5326	0.5241
(80:20)	AdaBoost	53.92 %.	0.5477	0.5392	0.5362

### **Approach-1: Evaluation**

#### **Confusion Matrix**







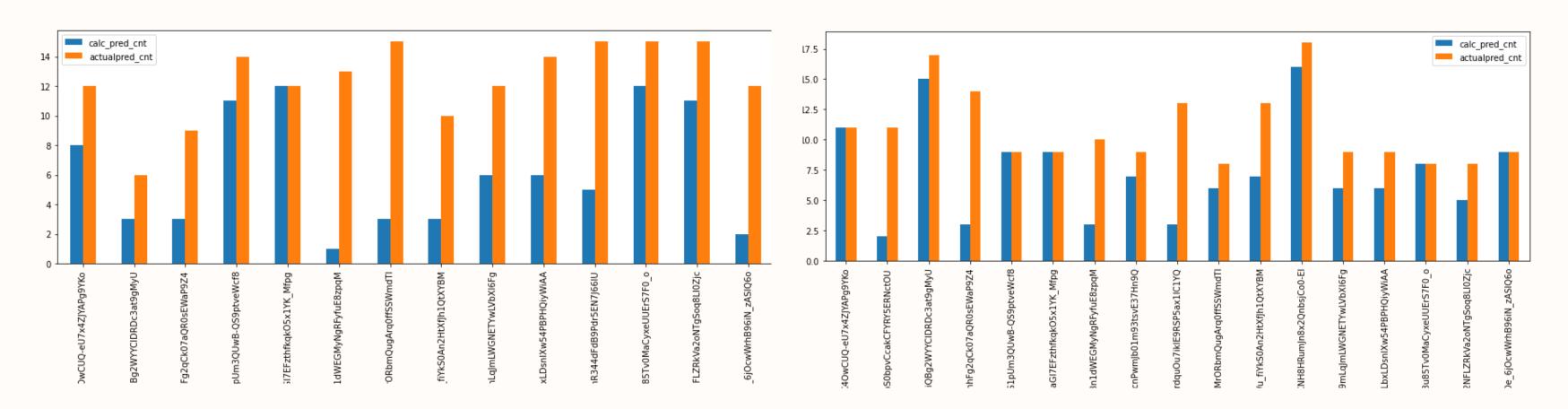


#### How soon can the onset be detected



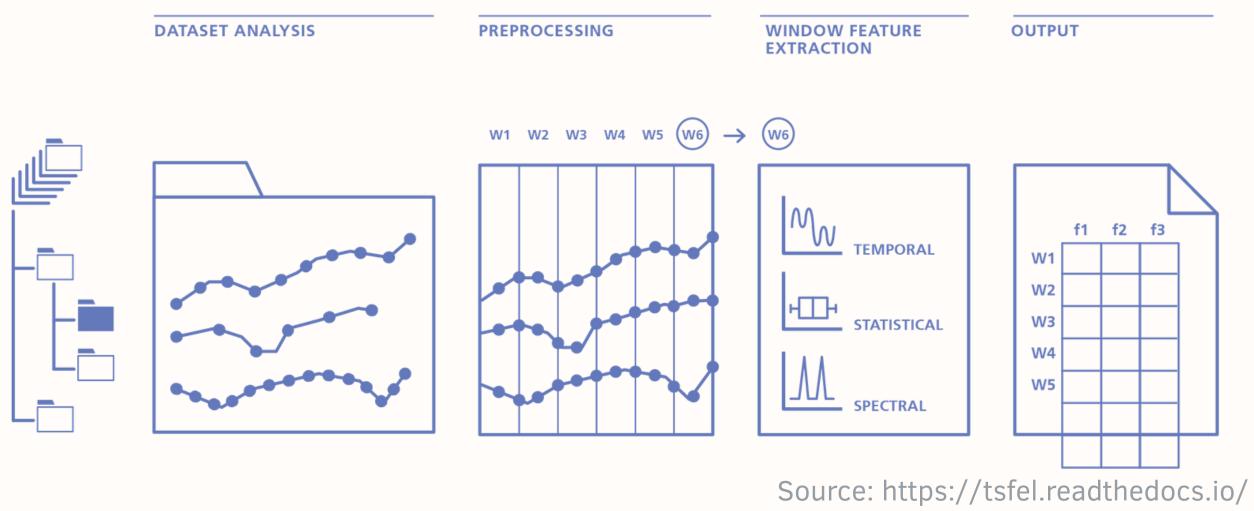
#### **Approach-1: Evaluation**

Analysis True Positives (Day Level per PID)



#### Analysis False Positives (Day Level per PID)

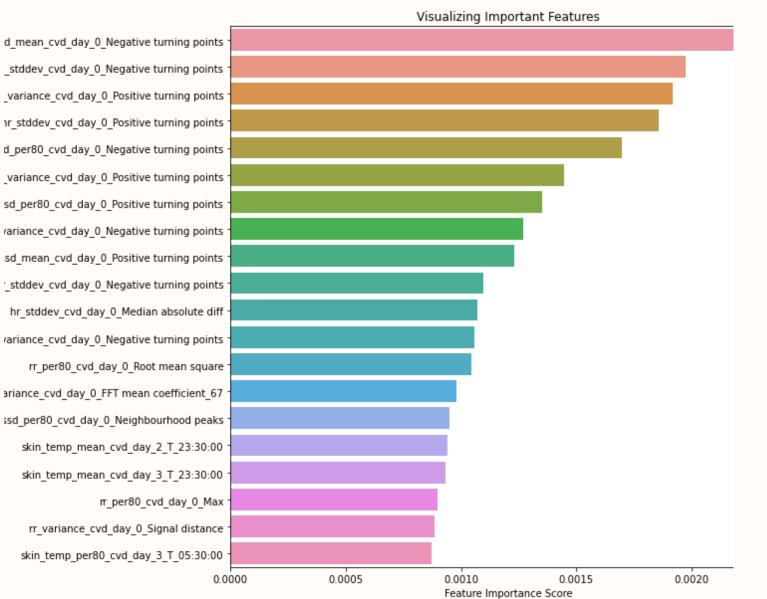
#### Approach-2: Time Series Feature Extraction Library



Input data shape : 1530 \* 11388 Total Population : 90 Train Test split : 80/20

#### Approach-2: Feature Importance & Results

#### Feature Importance

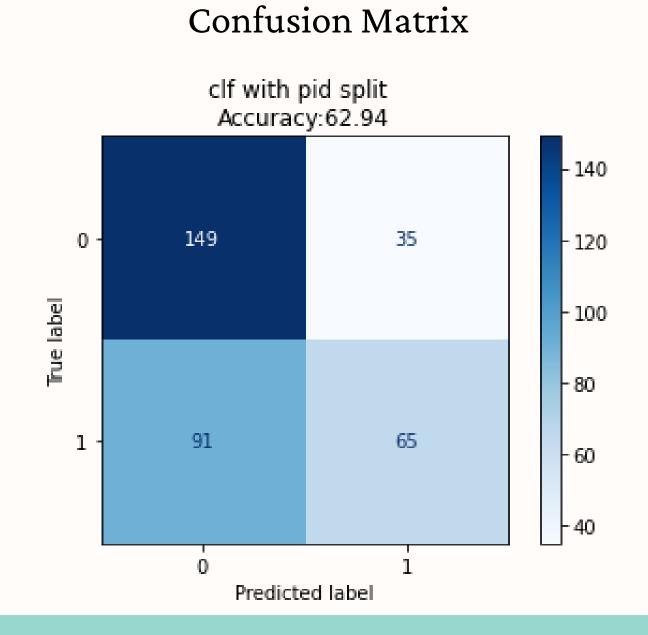


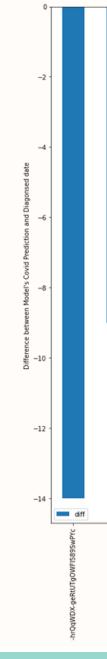
Test train PID split	Model/Classifier	Accuracy	Precision	Recall	Fscore
(80:20)	RandomForest	58.82 %	0.6051	0.5882	0.5802
(80:20)	XGB	53.59 %	0.5893	0.5784	0.5752
(80:20)	Bagging	58.17 %	0.5991	0.5816	0.5727
(80:20)	GradientBoosting	57.84 %.	0.5843	0.5784	0.5768
(80:20)	AdaBoost	58.5 %	0.5897	0.5849	0.584



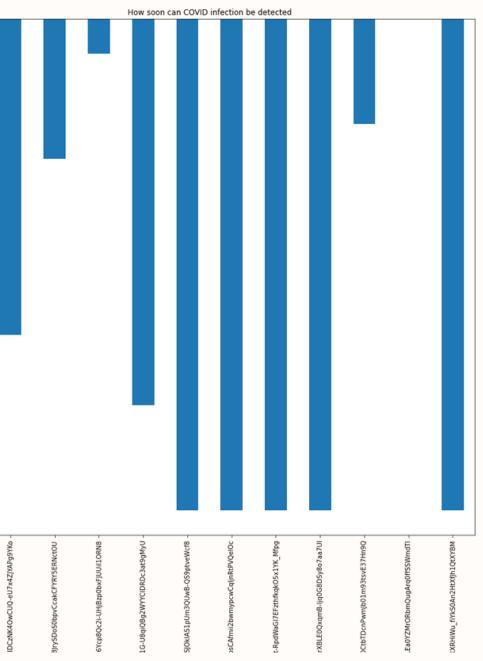
#### Comparison between different algorithms

### **Approach-2: Evaluation**



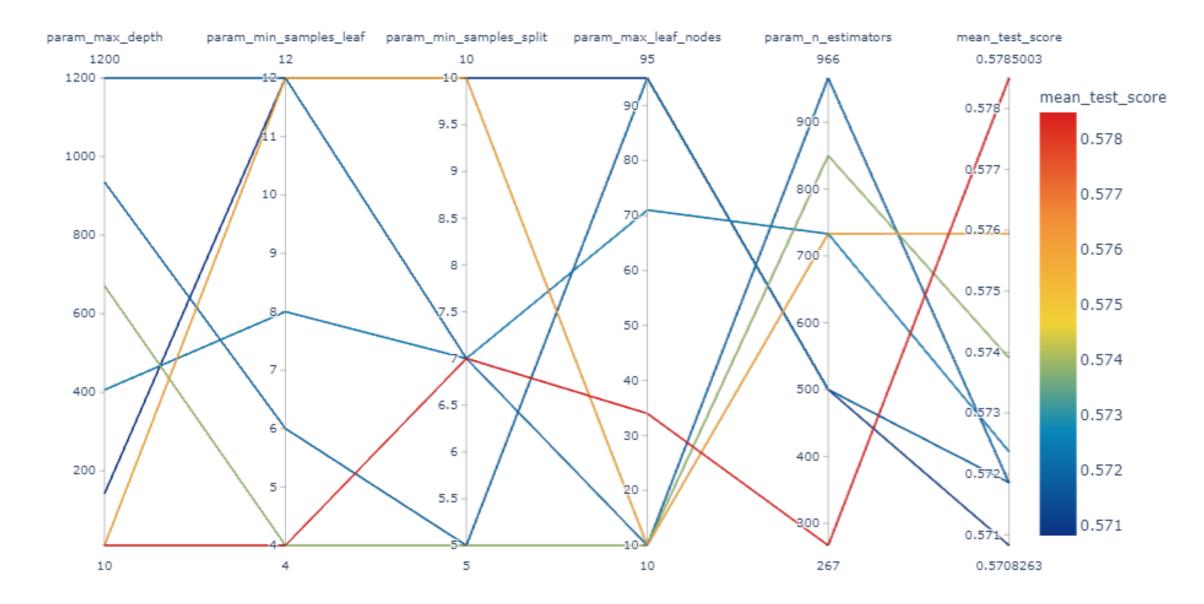


#### How soon can the onset be detected



### **Approach-2: Evaluation**

#### Parallel Coordinate Plot to select best parameters

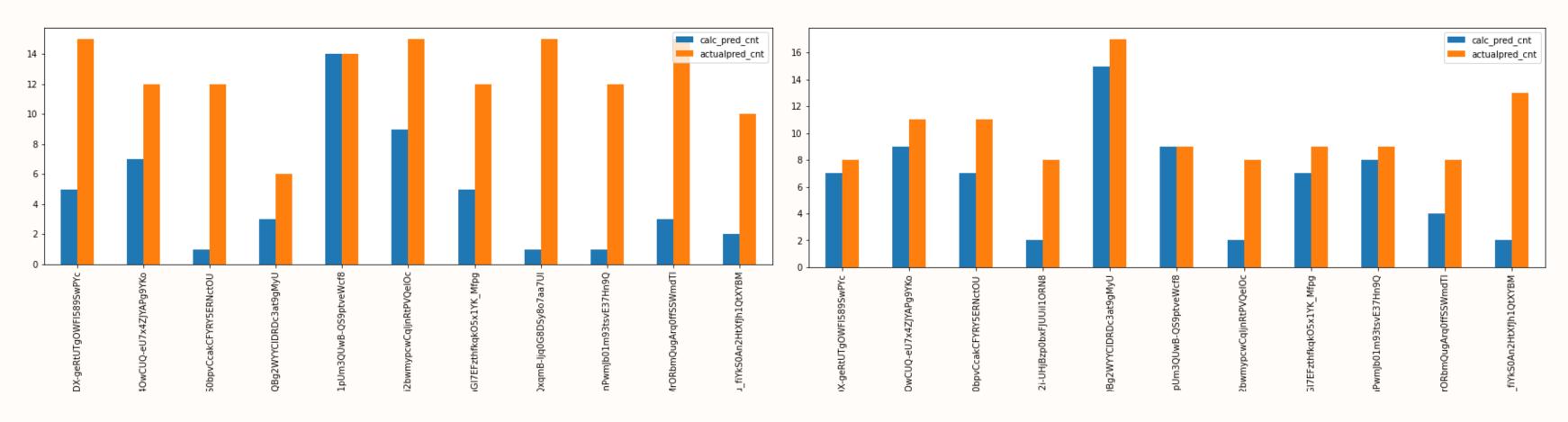




#### **Approach-2: Evaluation**

Analysis True Positives (Day Level per PID)





#### Analysis False Positives (Day Level per PID)

### **Approach-3: Aggregated Baseline Statistics**

Append the day based baseline's statistical feature

Higher order features Ratio of temp & met Ratio of HR & RMSSD

**Deviation from Baseline** 

Time Series Signal

**Original Data** 

x1

x1

x1

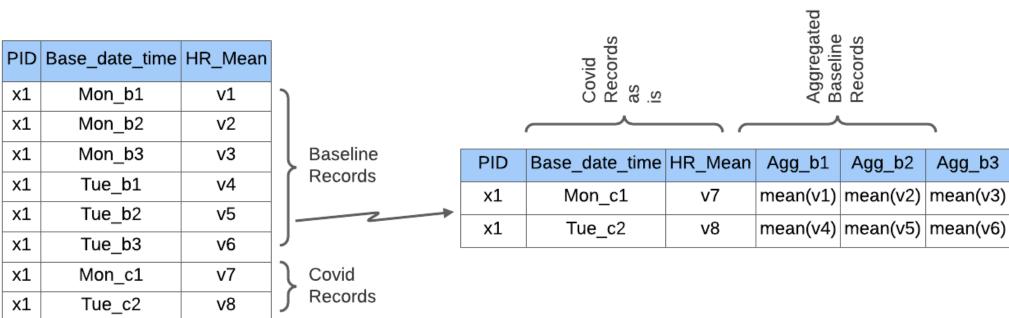
x1

x1

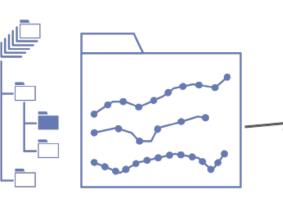
x1

x1

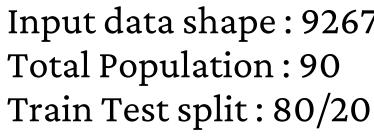
x1



**Baseline Aggregated** 



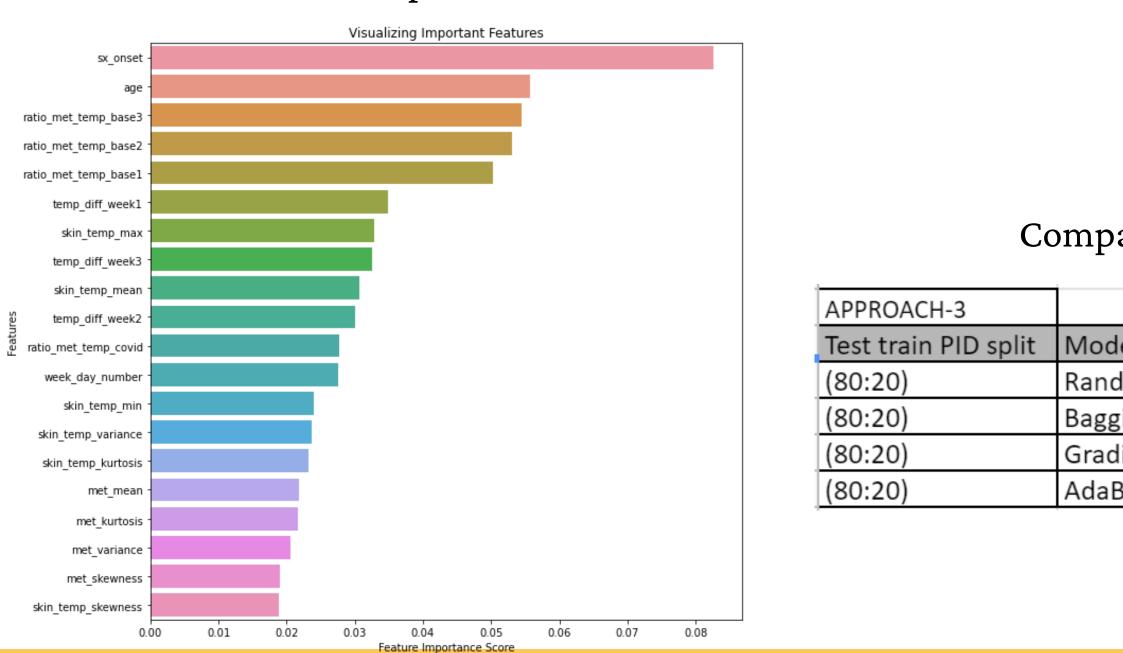




## Input data shape : 92671, 60



#### Approach-3: Feature Importance & Results



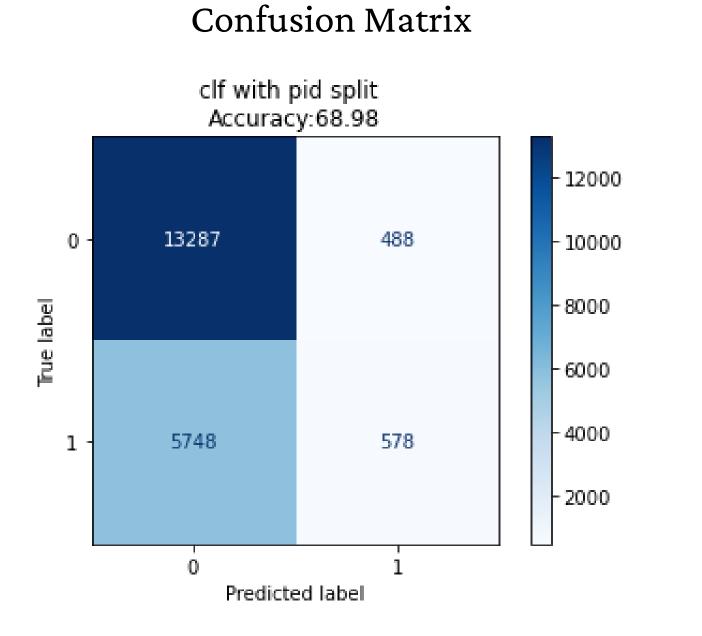
#### Feature Importance

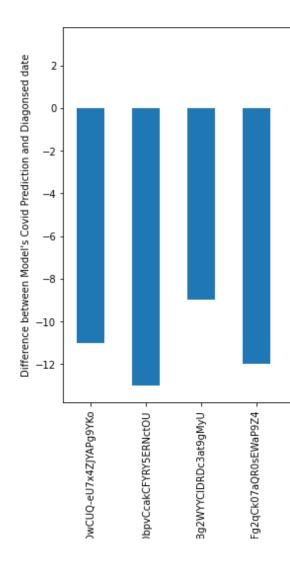
#### Comparison between different algorithms

del/Classifier	Accuracy	Precision	Recall	Fscore
domForest	69.17%	0.6626	0.6915	0.5936
ging	68.94 %	0.6531	0.6894	0.5889
dientBoosting	55.79 %.	0.5826	0.5578	0.5679
Boost	65.84 %	0.6041	0.6584	0.6091
Boost	65.84 %	0.6041	0.6584	0.6091



#### **Approach-3: Evaluation**







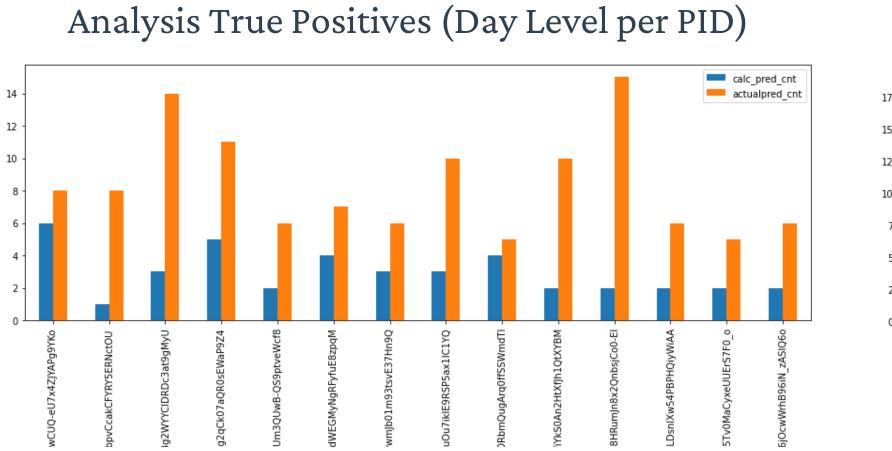
#### How soon can the onset be detected

diff JOU7ikIE9RSP5ax1IC1YQ mJn8x2QnbsjCo0-EI 35Tv0MaCyxeUUErS7F0\_0 Jb01m93tsvE37Hn9Q iiYkS0An2HtXfJh1QtXYBM R344dFdB9Pdr5EN7J66II dWEGMyNgRFyfuE8zpg Jm3QUwB-QS9ptveWcl IXw54PBPHQiyWi nQugArq0ffSSWr 6jOcwWrhB96iN

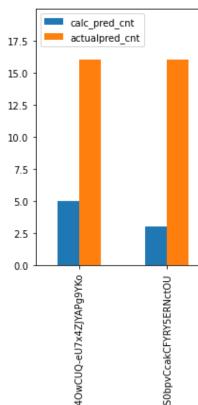
How soon can COVID infection be detected

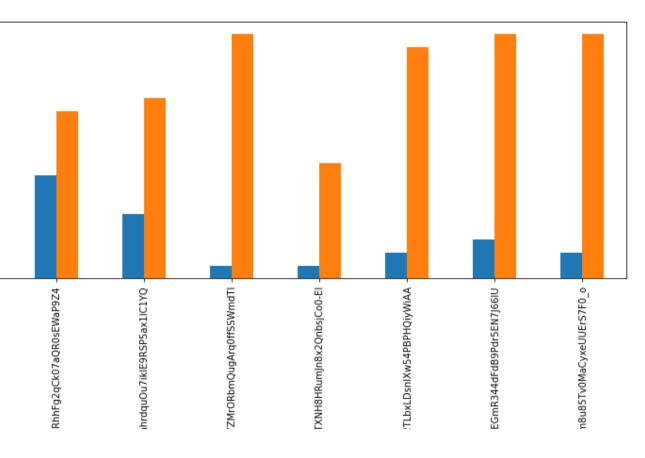


#### **Approach-3: Evaluation**



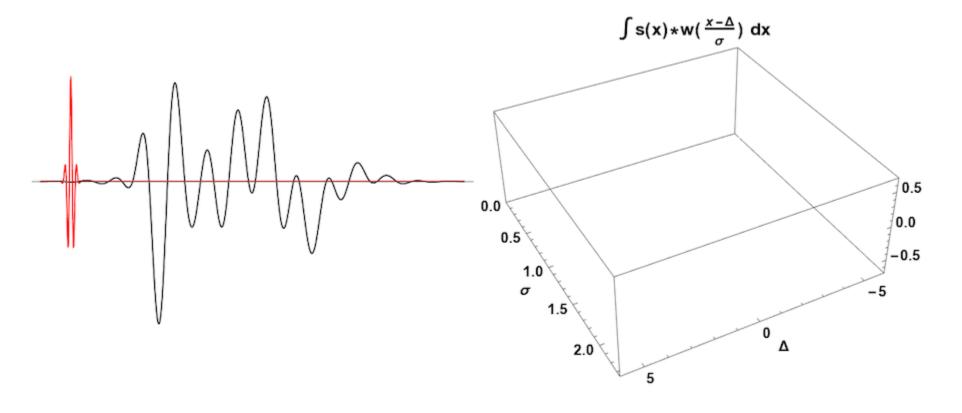
#### Analysis False Positives (Day Level per PID)





#### Approach-4: Pattern Recognition in Frequency Domain

Continuous Wavelet Transform (CWT)

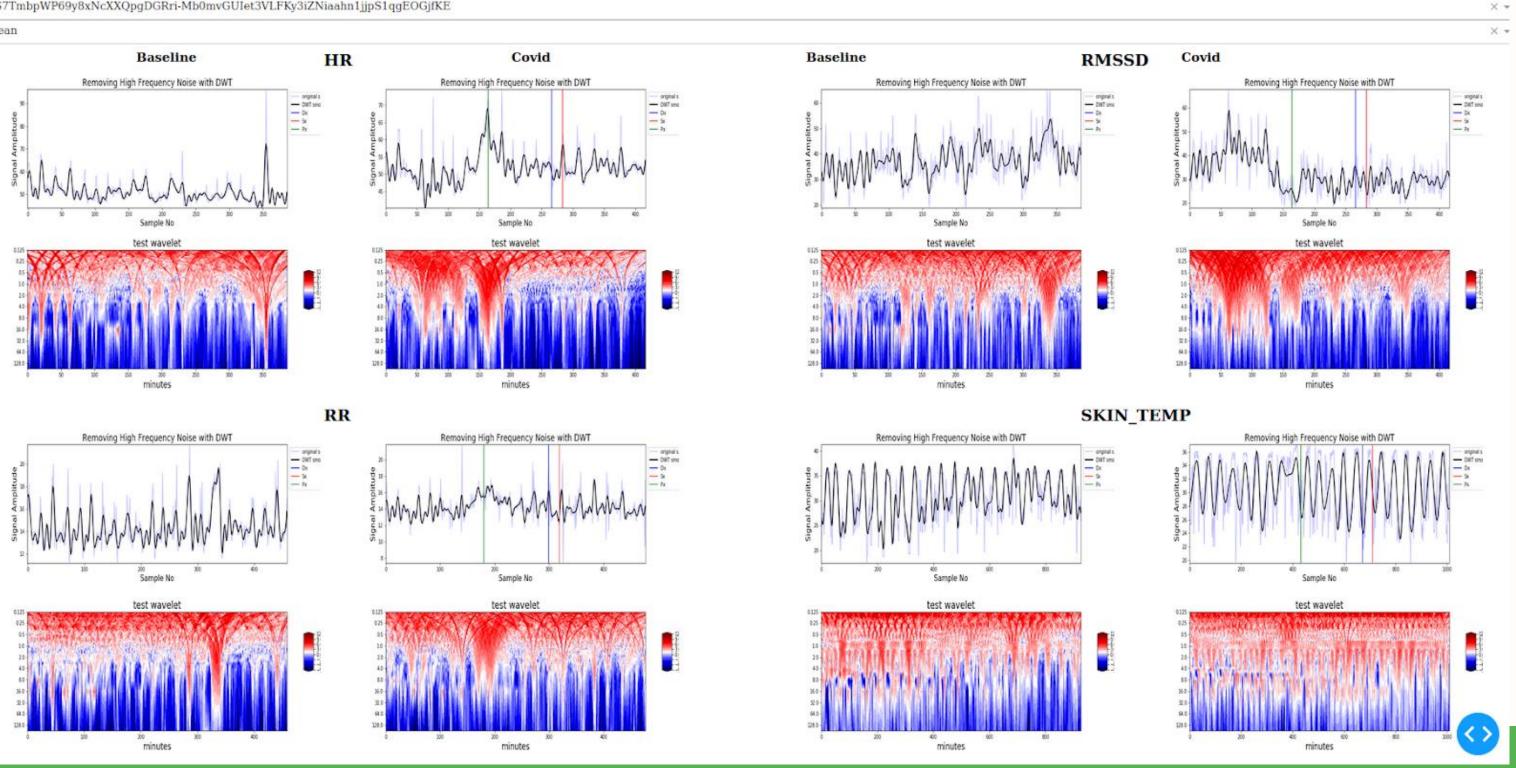


Source: wikipedia

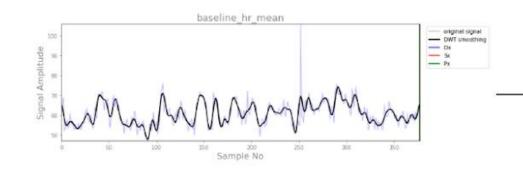
### Approach-4: EDA - CWT

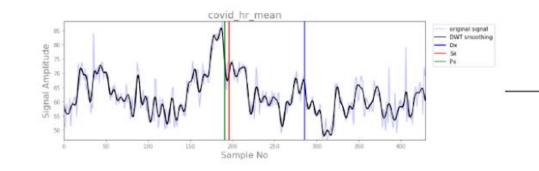
YS7TmbpWP69y8xNcXXQpgDGRri-Mb0mvGUIet3VLFKy3iZNiaahn1jjpS1qgEOGjfKE

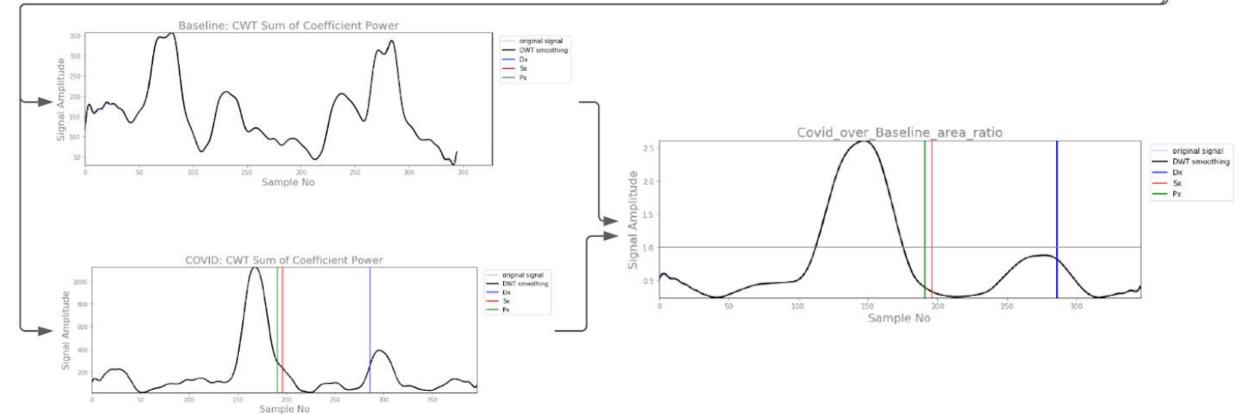
mean

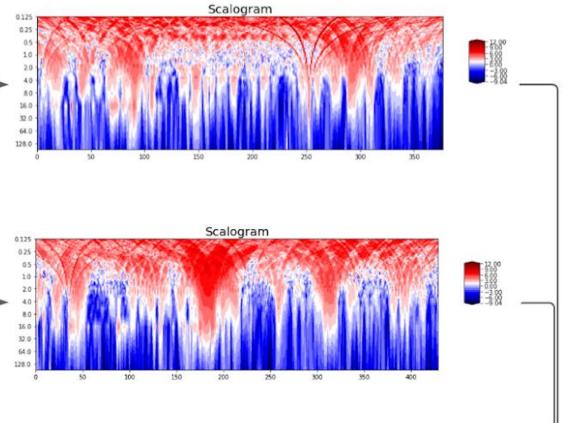


### Approach-4: CWT- Model









#### **Approach-4: CWT- Output Interpretation**

hr mean skin temp kurtos 0 10 20 30 40 50 60

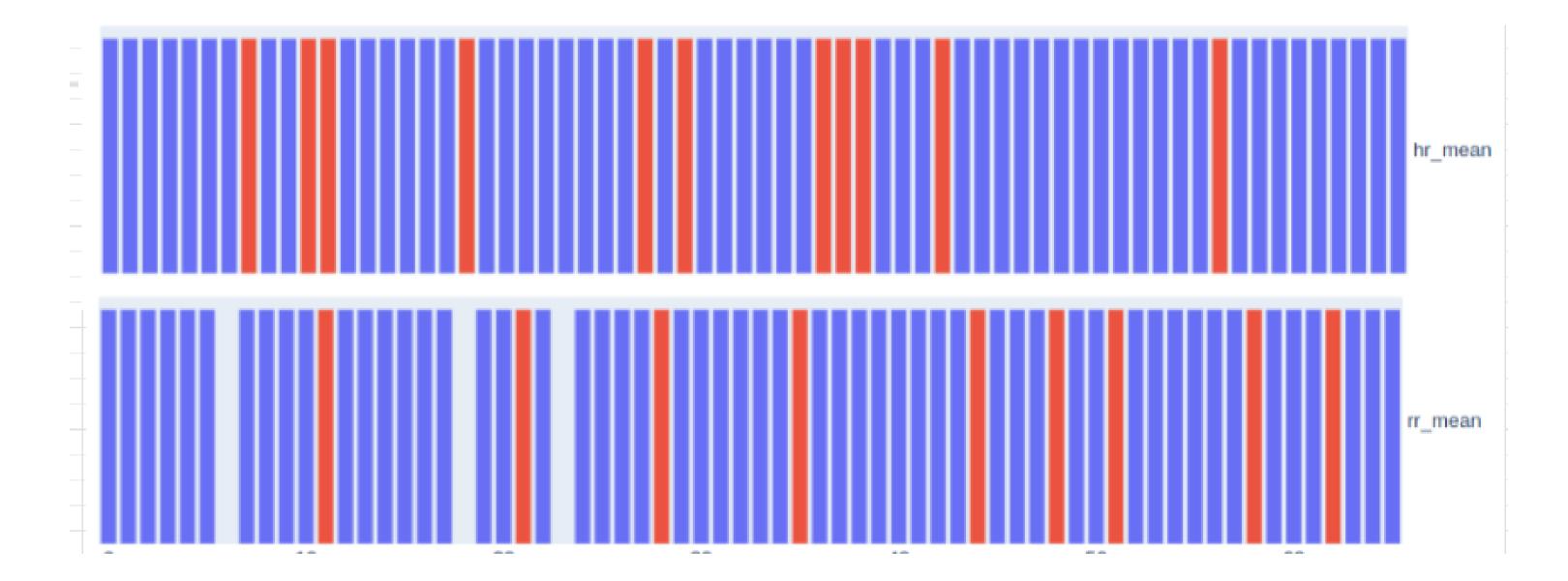
hr variance hr skewness hr kurtosis rmssd variance rmssd skewness rmssd kurtosis rmssd mean rr mean rr variance rr skewness Image: Second met mean met variance met skewness met kurtosis skin temp mean 💶 💶 🔳 skin temp varian skin temp skewr

#### ratio\_greater\_than\_one

True

False

#### Approach-4: CWT- Parameter Selection



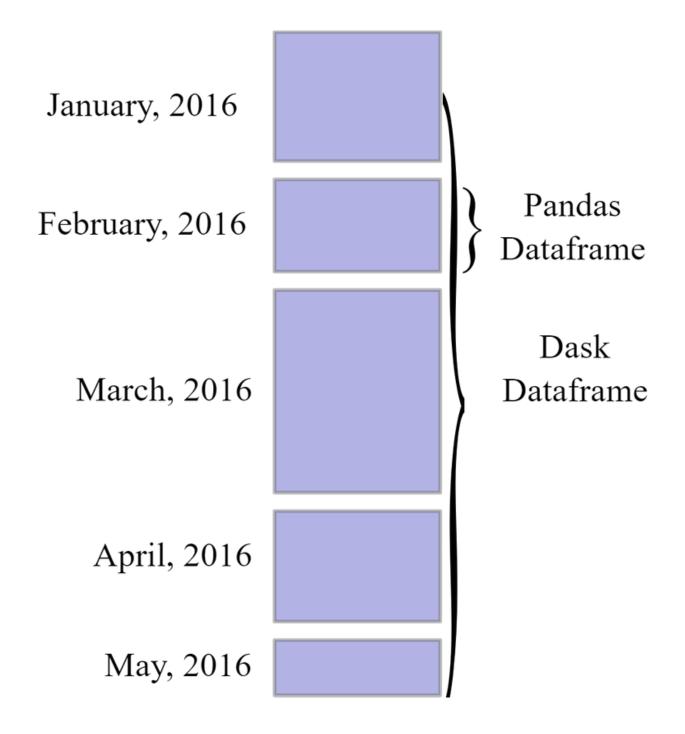
### Scalability



## Scalability

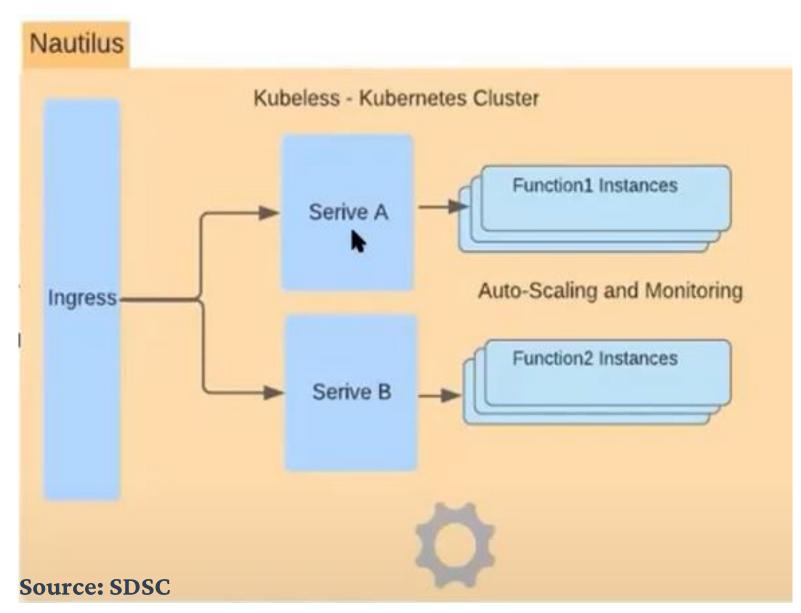
### Dask

- Dask is a flexible library for parallel computing in Python
- Dask DataFrame is a large parallel DataFrame composed of many smaller Pandas DataFrames.
- Dask-DataFrames may live on disk for a single machine, or on many different machines in a cluster.
- Dask DataFrame uses the multi-threaded scheduler that exposes parallelism

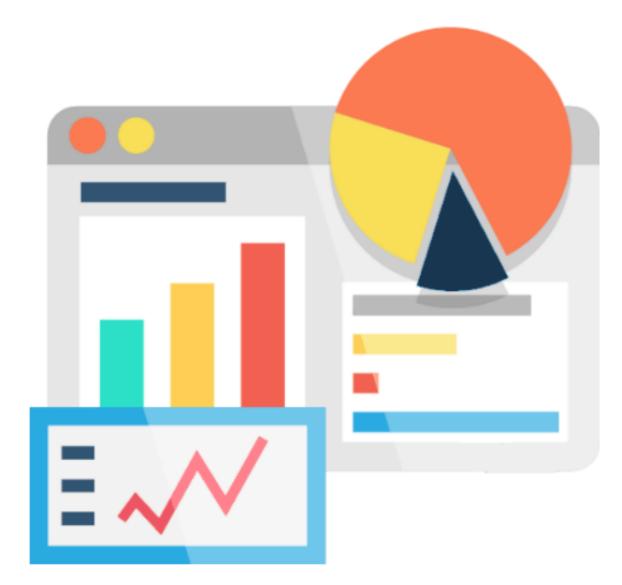


### Scalability

- Nautilus environment is built with Kubeless
  Kubernetes cluster architecture.
- Leveraging the existing auto-scalable infrastructure.
- When an instance is loaded beyond 80% by a service, a new instance will be automatically spun, and the load will be shared with the new instance.



### Visualization



### Visualization

Tools & Techniques

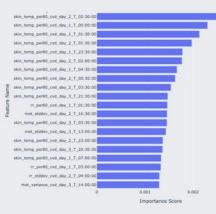
- Load Pickles from nautilus
- Ploty plots
- Plots integrated with DASH
- HTML,CSS and Dash-Bootstrap

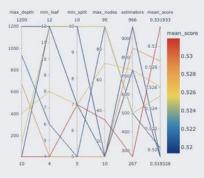




# 

#### IoT Wearable Data Fever Analysis Full View ML Approach 2 ML Approach Approach1 (Transposed Data 1190/340 10516 90.00% Covid Prediction (Person) 70/20 60.59% Train PID Count/ Test PID C Model Input Format Parallel Coordinate Plot to select best paramete





### Visualization

#### Setup & Extensibility

- Virtual Environment
  - Why
  - How

Easy to Extend for more approaches

• Easy to add more plots

ML Approach	Data Fever Analysis	S		Full V
Exploratory Analyzing the physiologic	Data Analaysis		dy the daily and weekly rhythm of each indiv	idual. The plots helps to compar
Time Period is from 2020-01-16		Choose Baseline/Covid All Baseline Covid	Choose a Label All O Covid Only	
Data Analysis ML Approach	ach 2 ML Approach 3			Full View
	h1 (Transposed Data		d onset. Raw data at the 30 minute interval in bo	oth the baseline and covid are
60.59% Model Accuracy	<b>1190 / 340</b> Train / Test Split	10516 Feature Count	<b>70 / 20</b> Train PID Count/ Test PID Count	90.00% Covid Prediction (Person)

F	u	Vi	e	w
-	<b>u</b>	¥1	6	

### Demo

2021-06-03



### Key Findings & Future Use

### **Key Findings**

- Skin temperature is the key feature predicting Covid in the time domain
- Heart rate and Respiratory rate are important features in the frequency domain
- The model performs better when all three variables are combined.
- Skin temperatures are higher for female than male (healthy and covid window)
- Higher-order features of baseline are the key features in the baseline aggregation approach Ratio : skin temperature / met, HR / RMSSD Deviation: the difference between covid to corresponding three baselines

### **Key Findings**

- Baseline Data was helpful in reducing false positive results
- TSFEL approach signifies entropy (amount of uncertainty) and negative turning points are the important features for the predictions
- Most of the models' true prediction is around the Symptoms onset date or the calculated px date (Calculated physiological max date)

### **Future Use**

- The framework allows easy integration of new physiological data. Ex. Sleep Summary
- The framework facilitates the development and evaluation of new model approaches or selecting dynamic baseline
- Extend the research for other Medical diagnosis like pregnancy prediction
- Ease of collaboration among different stakeholders (for eg. clinicians, algorithm developers & physicians)



### Acknowledgment

Prof. Ilkay Altintas	Shweta Purawa
Program Advisor	Computational
Prof. Benjamin Smarr	Paul Norton
Academic Advisor	Security special
Joseph Natale	Danielle Whitel
postdoctoral researcher	Technical proje
Subhasis Dasgupta	Shakti Davis
Computational and Data Science Researcher Specialist level 4	Engineer, MIT L
Varun Viewanath	

Varun Viswanath Ph.D. student, Dept. of Electrical and Computer Engineering at al and Data Science Researcher Specialist level 4

list, San Diego Supercomputer Center

ehair ect manager, San Diego Supercomputer Center

Lincoln Lab





