SAND2019-14439C

iPredictome: A Data Science Toolkit for Predictive Modeling for Health Anomalies

PRESENTED BY

Kunal Poorey

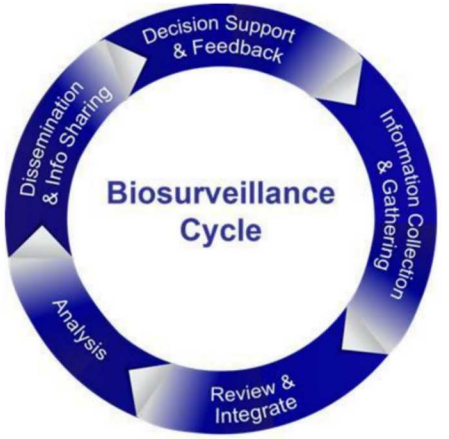
Senior Member of Technical Staff

Sandia National Labs, Livermore, CA

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

Importance and Need

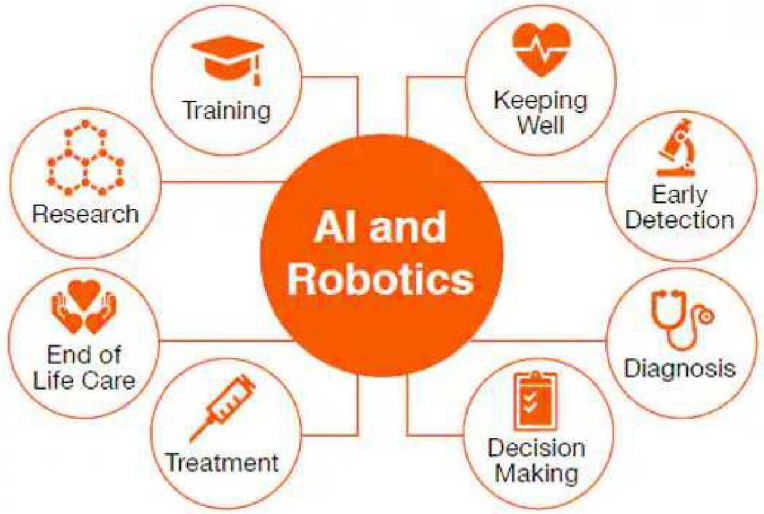
- With the growth of population and increasing globalization combined with emergence of antibiotic resistance in pathogens the potential threat of global pandemic is on the rise.
- Emerging new pathogens makes it difficult to detect an outbreak early on.
- Current state-of-the-art Biosurveillance strategies are disconnected, sparsely applied, and confined to few agents
- No clear strategy for individual early warning health monitoring
- Globalization and United states' interest in maintaining global peace, US troops are deployed all over the world. These unfamiliar environments pose exposure to new pathogens to US warfighters
- With the advancements of Data Science and Biological research , assays and sensors we are now exposed to advanced tools relevant for effective early prediction of an "event"



Data Science role in Health Informatics

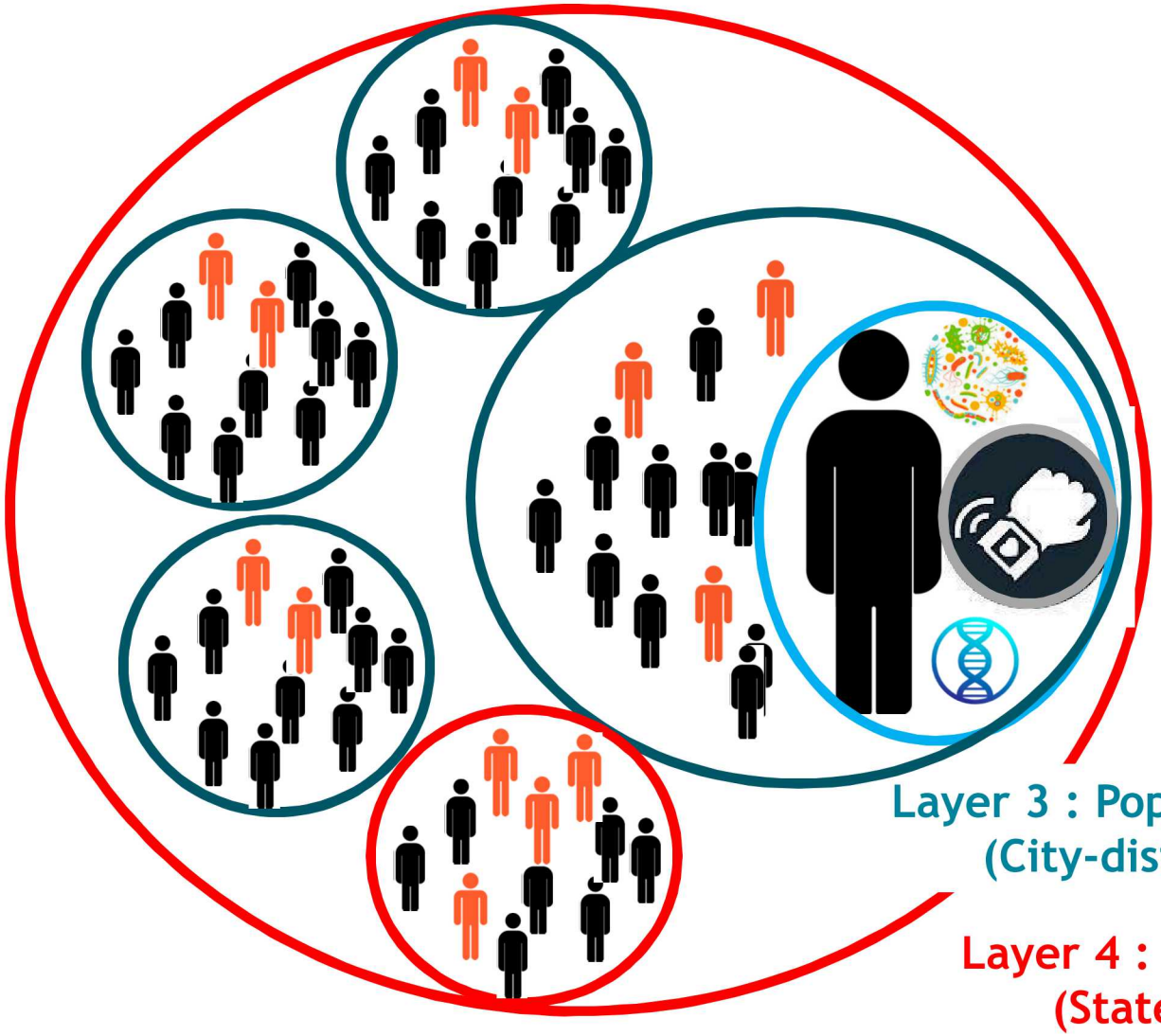
- Data science applications have made feasible to understand complex biological processes
- Better signal/noise, find patterns in large datasets and improving accuracy, efficiency and early detection
- Application of machine learning and AI has been aiding in decision making process in medical diagnostics
- Advancement of precision and personalized medicine driving deposition of personalized genomics and metagenomics data

Data Science and machine learning is now an integral part to understand, characterize and detect patterns in biological assays including genomics/metagenomics, proteomics, cheminformatics, imaging, etc.



WHATS NEEDED ?

Layered detection of anomalies in health dataset



Layer 1 : Individual assay Individual Health Anomaly detection

Layer 2 : Individual Health Anomaly detection

Layer 3 : Population Health Anomaly detection (City-district level)

Layer 4 : Population Health Anomaly detection (State-Nation level)

Availability of Human and Environmental Datasets

Currently there the available datasets can be characterized into two categories

Self/Personalized monitoring	Monitored or can be at health care facility
On a continuous time series for various conditions ◦ **Hearth rate** ◦ **Oxygen content** ◦ Breathing ◦ Movement ◦ Blood pressure ◦ Glucose levels	per visit bases ◦ Biochemistry tests for biomarkers ◦ Genomics profiling ◦ **Microbiome monitoring** ◦ Proteomics monitoring ◦ Metabolic activity ◦ VOCs

We have focused our attention to Data Science application to Microbiome profiling for 1st layer ML application.

Examples of some wearable health devices. (1)—SensoTRACK ear sensor; (2)—Google Contact Lens; (3)—BioPatchTM; (4)—Smartwatch Basis PEAKTM; (5)—QardioCore; (6)—Vital Jacket® t-shirt; (7)—Moov (activity tracker) Dias et.al 2018

Data Science Application for Disease Prediction

Microbiome and Human Diseases

Microbiome: Humans are 50-60% microbial cells (100 trillion cells). Microbiome is helps us carry out major biological functions as **digestion, immune response**

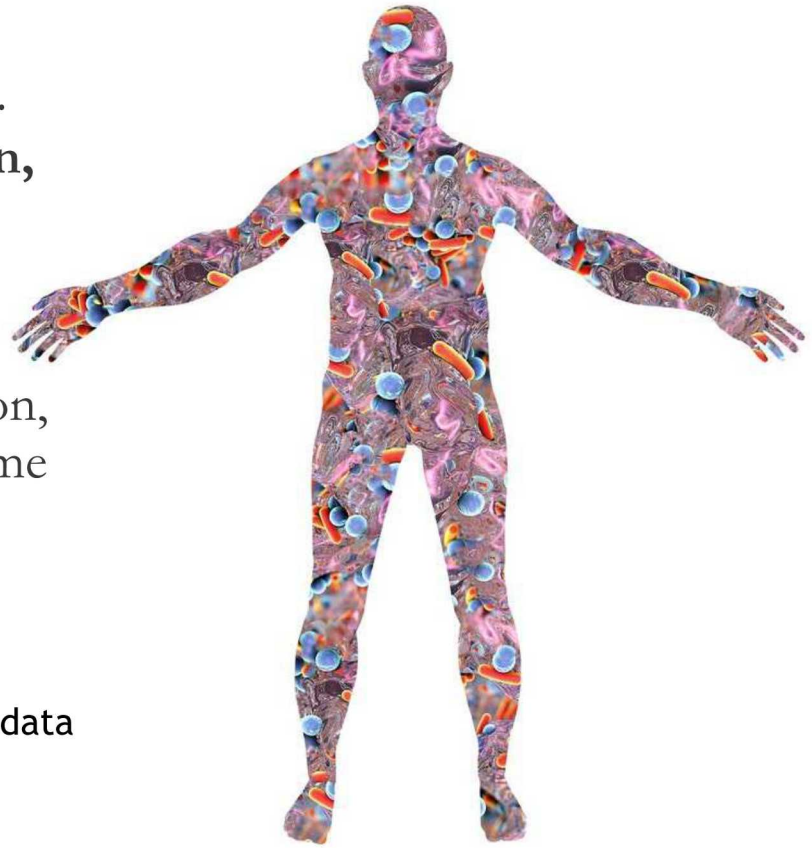
Humans are 50-60% microbial cells (100 trillion cells)

Microbiome is helps us carry out major biological functions as digestion, immune response, neural response. And dysbiosis of human microbiome has been associated with variety of health disease and disorders

Early prediction of diseases in Humans with Data Science tools

We applied a custom machine learning pipeline **iPredictome** to Human Microbiome Project data

1638 Samples and 572 variables

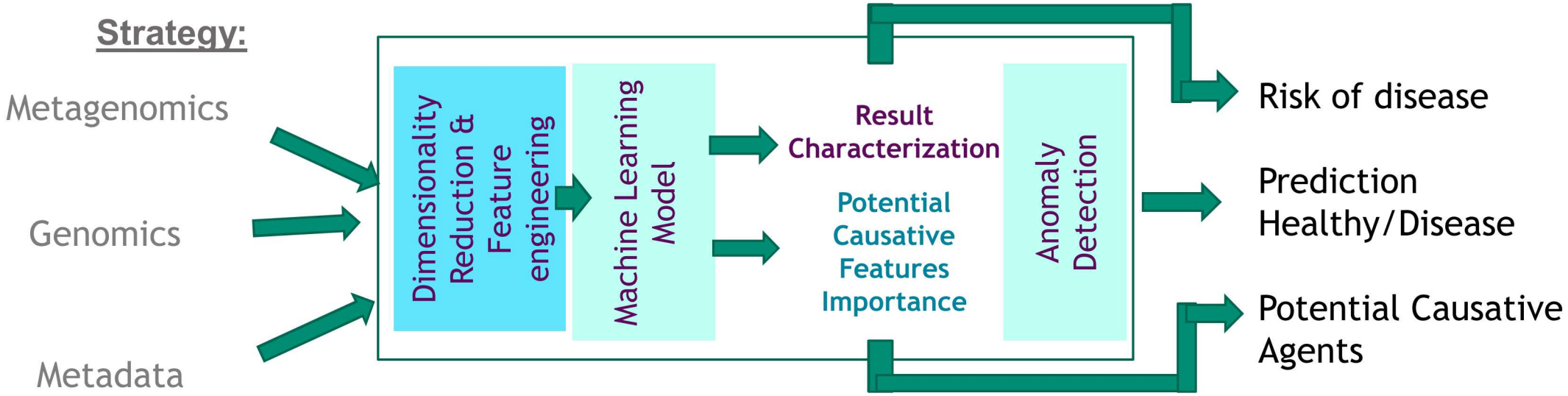


CREDIT: KATERYNA KON/SCIENCE PHOTO LIBRARY, VIA GETTY IMAGES

iPredictome

The iPredictome Strategy

iPredictome is the machine learning toolkit to detect anomalies in the Biological dataset to characterize the health state.



But also has potential to work with Proteomics, Metabolomics, VOCs, etc. and we are actively working to include them in our feature set .

iPredictome

Feature Engineering

Leanne Whitmore

iHMP NIH Integrative Human Microbiome Project

1638 Samples and 572 variables

16s ssu Amplicon Sequencing & Metagenomic Sequencing

Extract taxonomic profiles (metagenomic sequencing) and metadata from human microbiome project

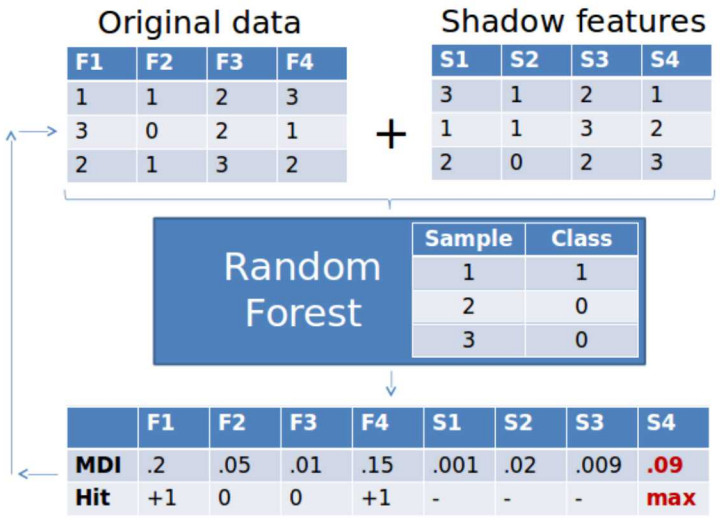
Use feature reduction algorithms such as Boruta and tSNE to initially evaluate important taxonomic profiles

Machine Learning
Test a number of classifiers to determine which best predicts each diseased state

Identify distinct taxonomic signatures using local interpretable model agnostic explanations (LIME)

Feature Reduction

Boruta : feature Selection with randomization



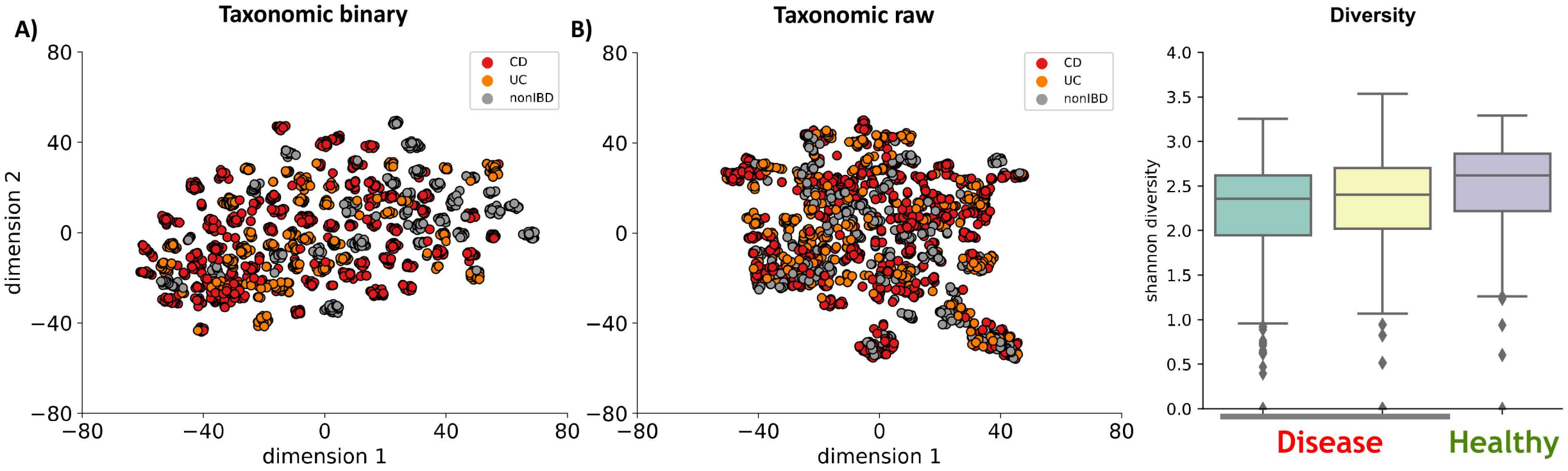
Model Explainer → Feature importance

LIME : Local Interpretable Model-Agnostic Explanations

iPredictome

Feature Engineering

Feature engineering for patterns in microbiome structure in disease



iPredictome

Performance

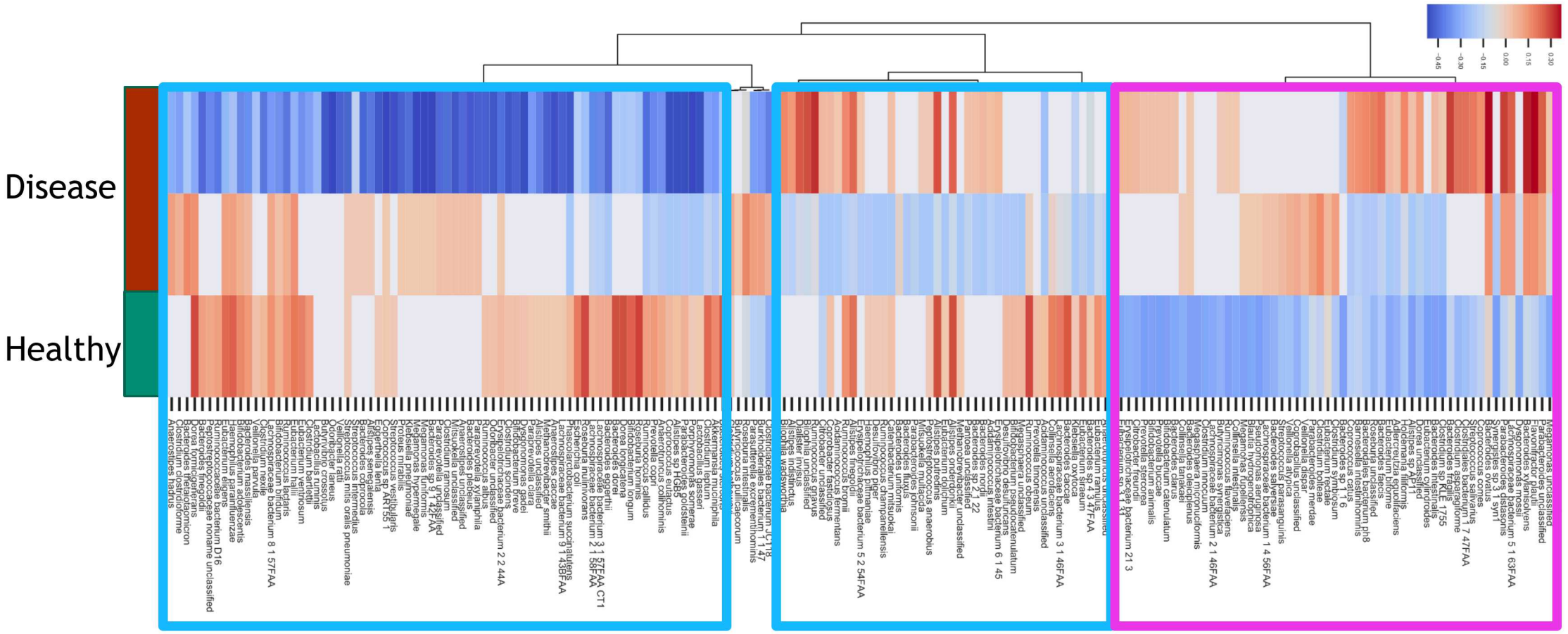
Classification Accuracies
Predictive Machine Learning Algorithm

Accuracy of iPredictome : > 96% accurate

Classifier	Raw full	Binary full	Raw reduced (boruta)	Binary reduced (boruta)
Kneighbors jaccard	0.96, (+/- 0.02)	0.96, (+/- 0.02)	0.95, (+/- 0.04)	0.96, (+/- 0.04)
Kneighbors euclidean	0.83, (+/- 0.06)	0.94, (+/- 0.04)	0.82, (+/- 0.06)	0.96, (+/- 0.04)
RandomForest	0.86, (+/- 0.06)	0.89, (+/- 0.04)	0.88, (+/- 0.04)	0.92, (+/- 0.04)
Multi-Layer Perceptron	0.5, (+/- 0.06)	0.89, (+/- 0.04)	0.53, (+/- 0.06)	0.9, (+/- 0.06)
AdaBoost	0.71, (+/- 0.06)	0.7, (+/- 0.04)	0.71, (+/- 0.1)	0.72, (+/- 0.08)
Support Vector Machines	0.45, (+/- 0.06)	0.67, (+/- 0.08)	0.44, (+/- 0.1)	0.8, (+/- 0.06)

iPredictome

LIME interpretation of ML Predictions



We observe significant more biomarkers correlated with diseases than originally reported

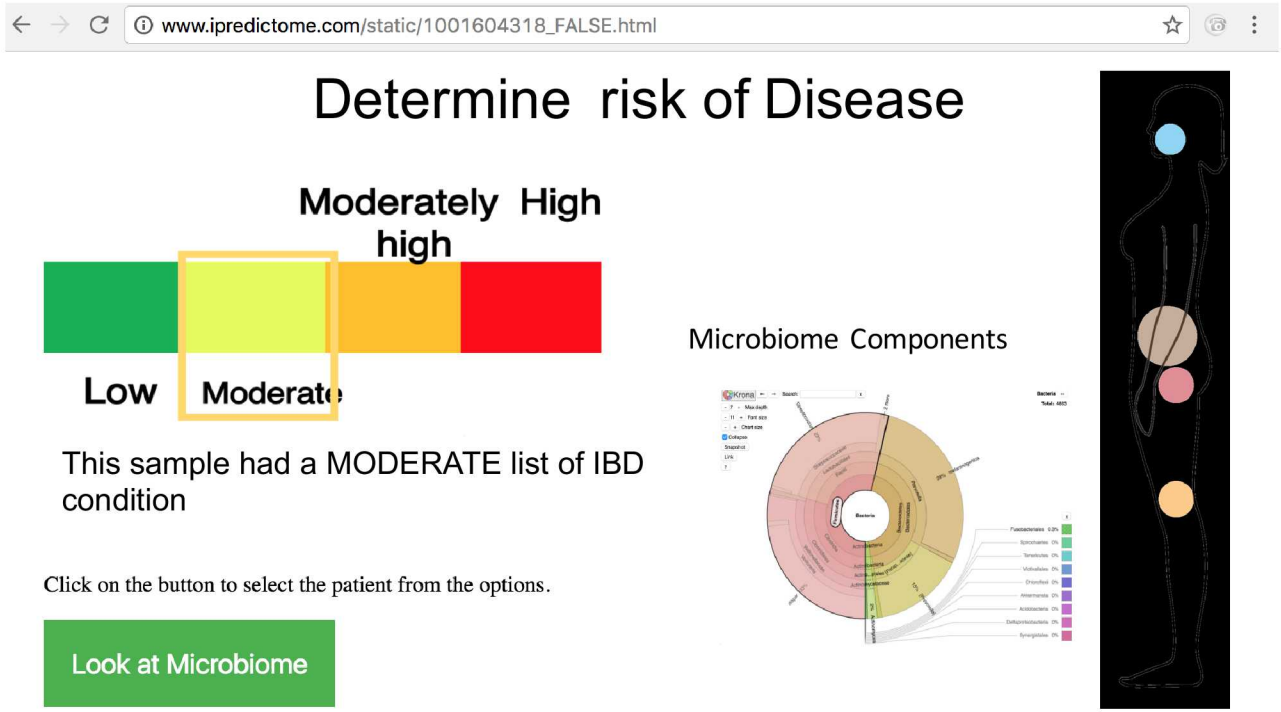
Additionally we can report patient specific biomarkers for potential personalized treatment

iPredictome

Future Application

Software has the capability to be hosted in cloud for prediction to be deliver through a web-app or mobile phone

An example output is displayed below.



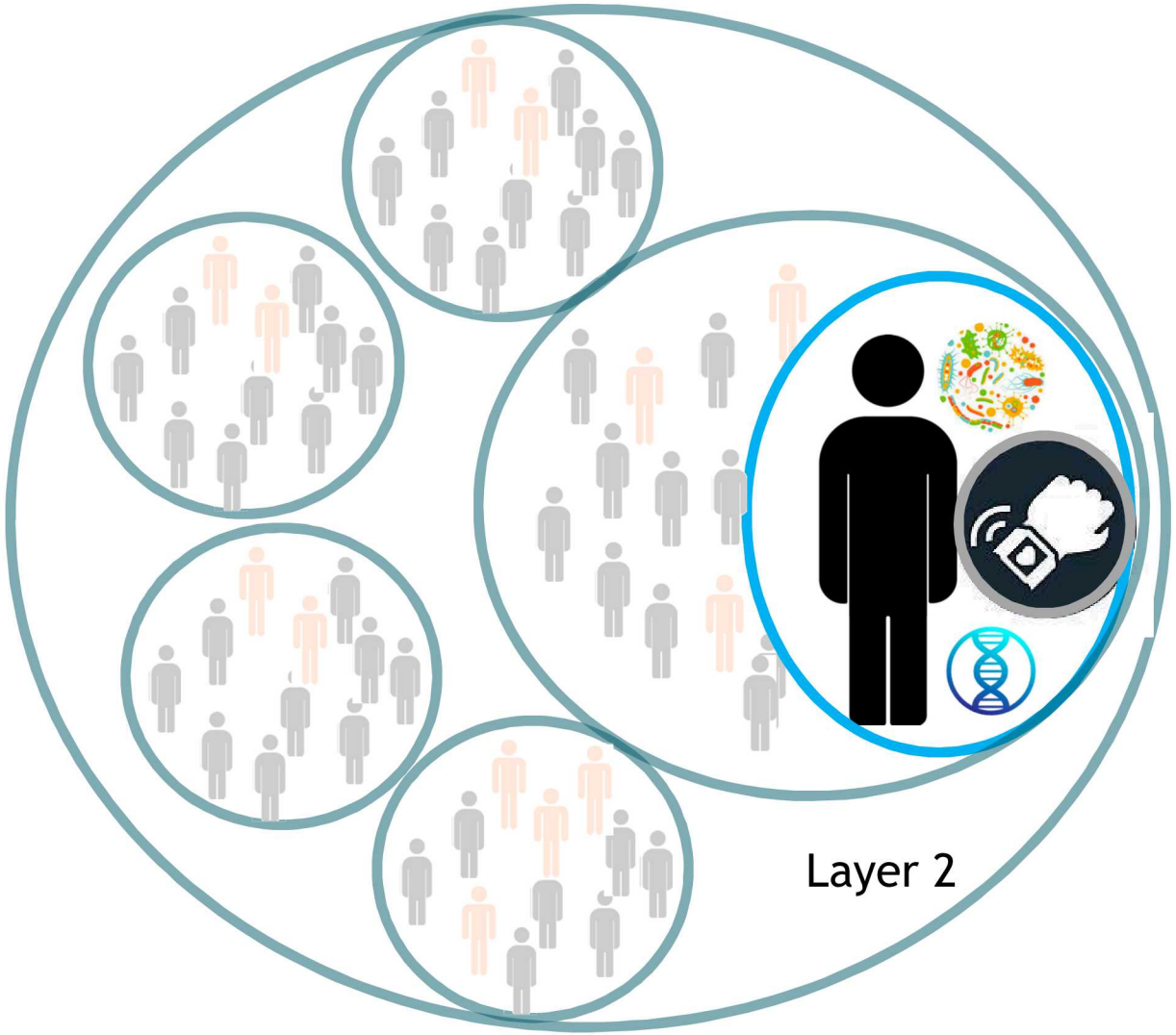
What's Next

Anomaly detection on the multiple data streams on Individual

We are attempting to apply Sandia developed method of unsupervised anomaly detection method to call out anomalies in wearable data.

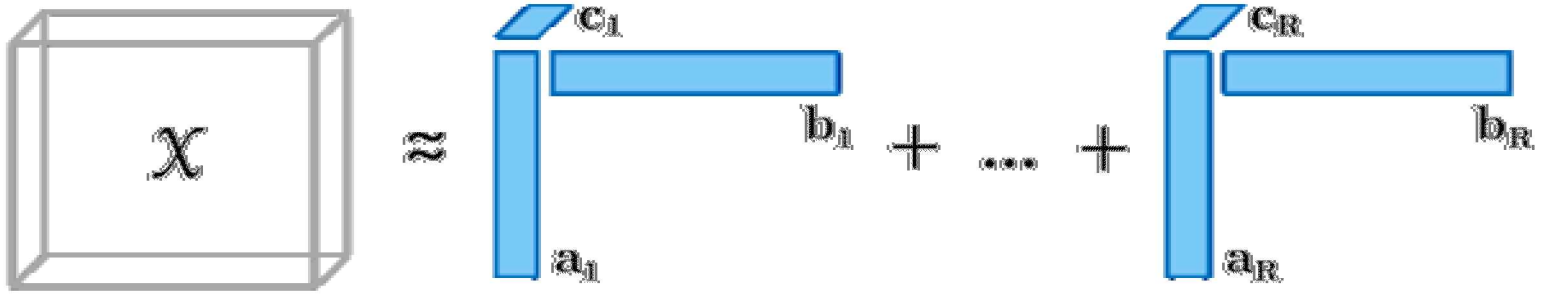
- Unsupervised Anomaly detection
- Based on Tensor Decomposition
- Ideal for multivariate dataset for detection of spatial or temporal variations

Konduri et al 2019 J. Comput. Physics



Anomaly detection detects changes in vital signs with flu like symptoms

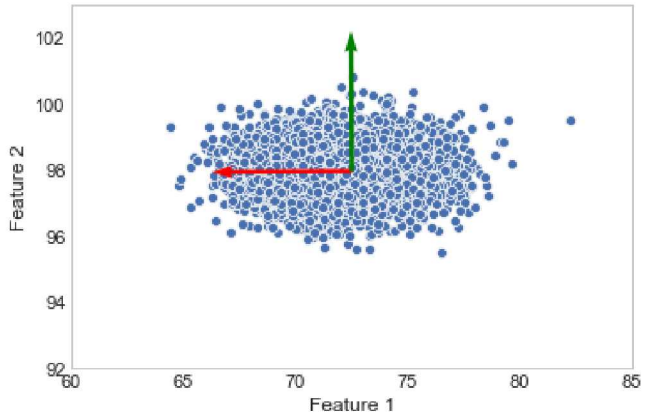
The algorithm decomposes the data into several spatial sub-domains and time steps

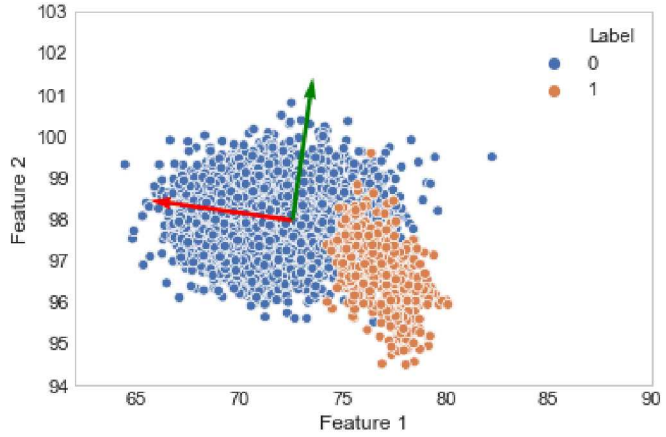


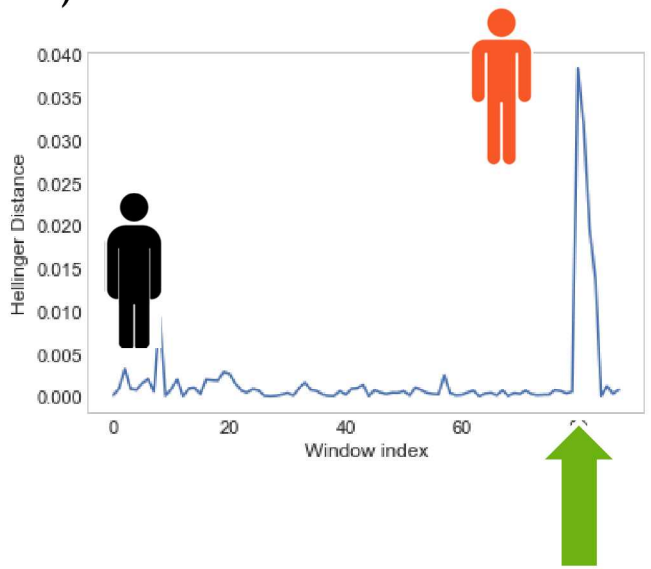
Hellinger distance

(Measures change in orientation of singular vectors weighted by singular values)

Anomaly event







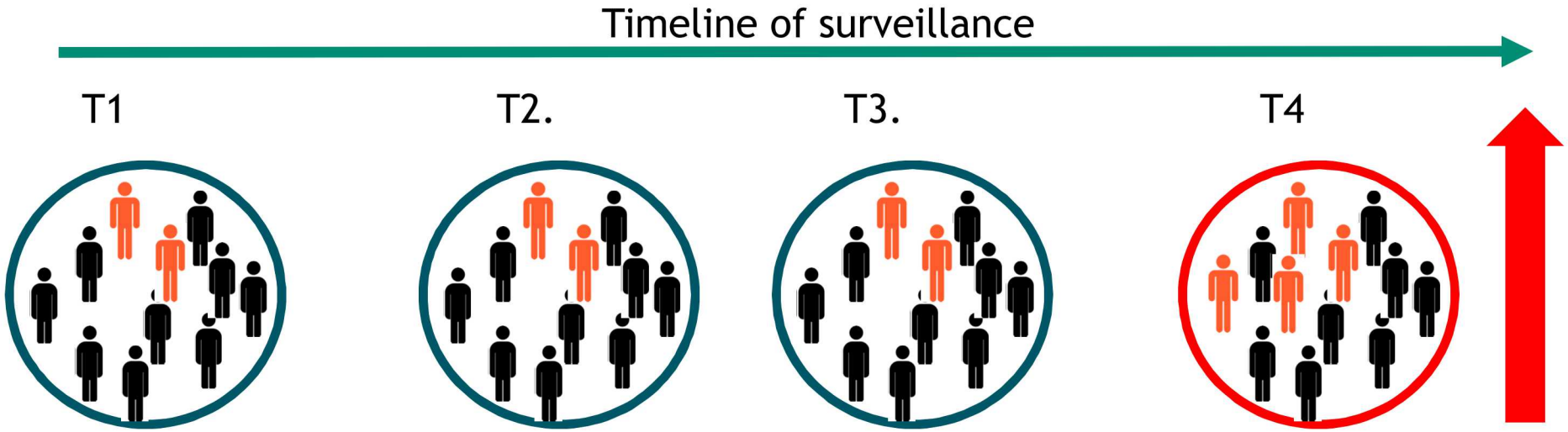
Anomaly detected from health data

Future Focus : Integration at higher levels

Anomaly detection on the population level

Turning the Individual Layer anomaly as a reading for the local population layer of detection for population level anomaly

Third Layer
Localized population anomaly detection



Anomaly in population health an indication of local outbreak

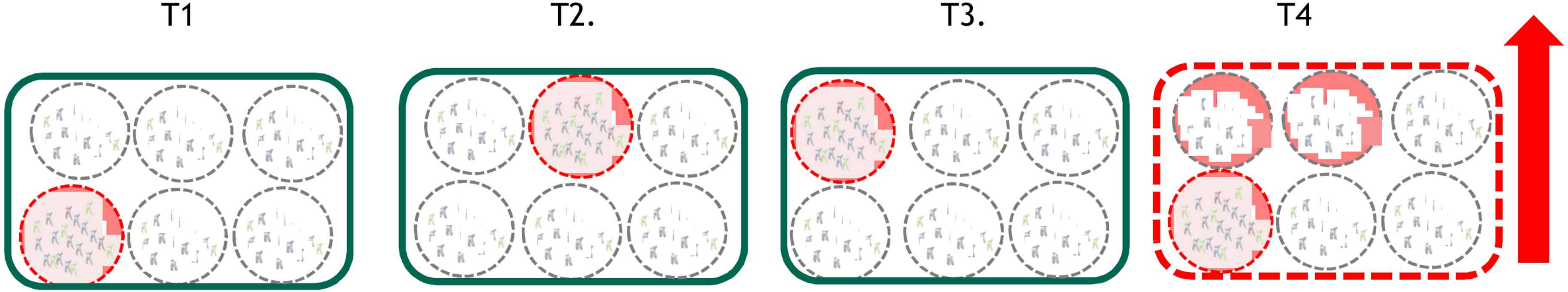
Future Focus : Integration at higher levels

Anomaly detection on the Global level

Turning the local population anomaly as a reading for the fourth layer of detection for Global level anomaly

Fourth Layer
Global population anomaly detection

Timeline of surveillance



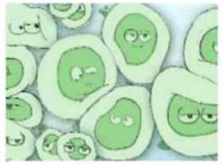
Anomaly in population health an indication of outbreak

Application of similar Data Science strategy for Crop Protection

Application of similar Data Science strategy for Crop Protection

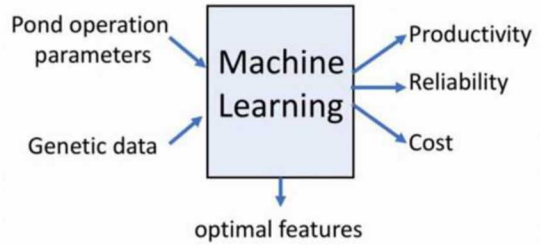
- Cost : for sustainable use we need **HIGHER** annual productivity
- Annual productivity Increase ways
  - Reducing biomass loss by **avoiding crashes**

Algae + **Agents: predators, pathogens virus & fungi** + **Environment** = **CRASH**



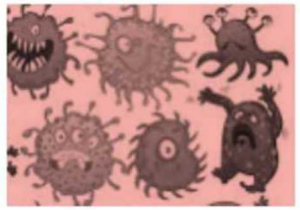

- Need of **higher daily productivity**

- **Annual Productivity**
  - **Reduce number of crashes.**
  - **Building Predictive Machine Learning Models for Crashes**



Good

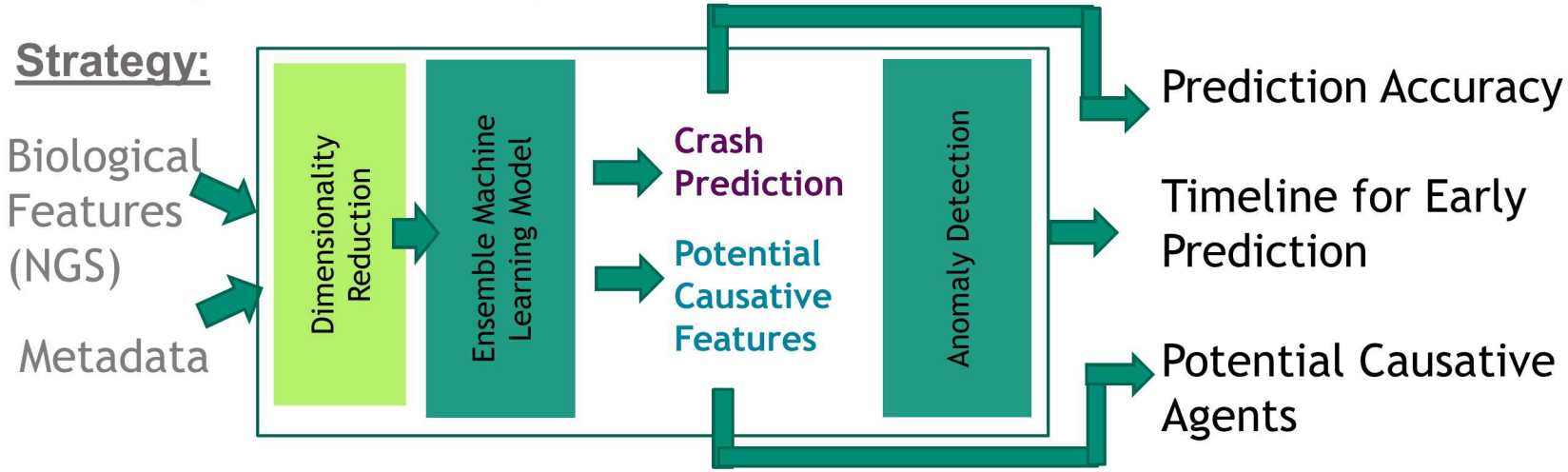
BAD



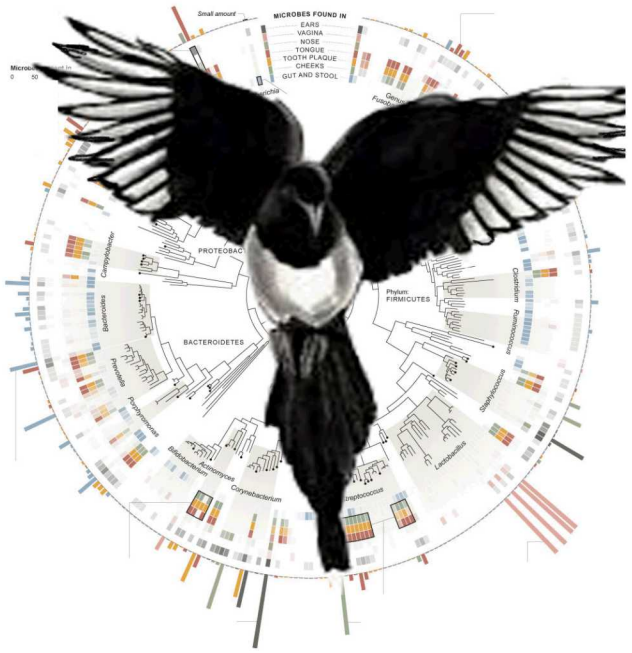
- **Increase daily productivity**
  - **Identify beneficial factors for productive growth**
  - **Identify deleterious factors**
  - **Identify conditions and strategy for optimized production**

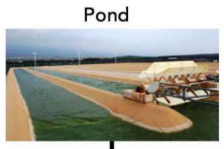
Build an algorithm for early detection of anomaly for each cultivation run

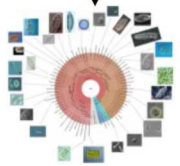
Strategy:

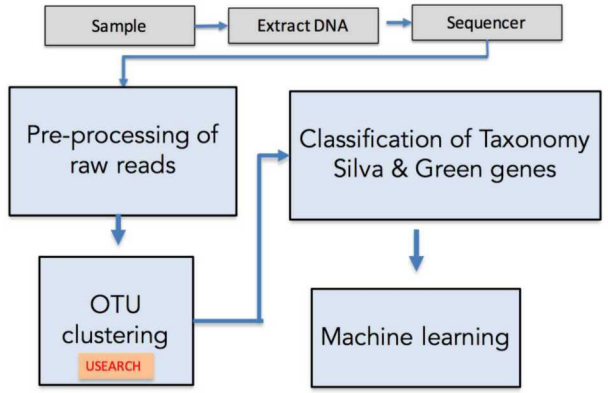


Application of similar Data Science strategy for Crop Protection









MAGPie Pipeline: A parallel processing amplicon sequencing analysis pipeline for ultra fast analysis

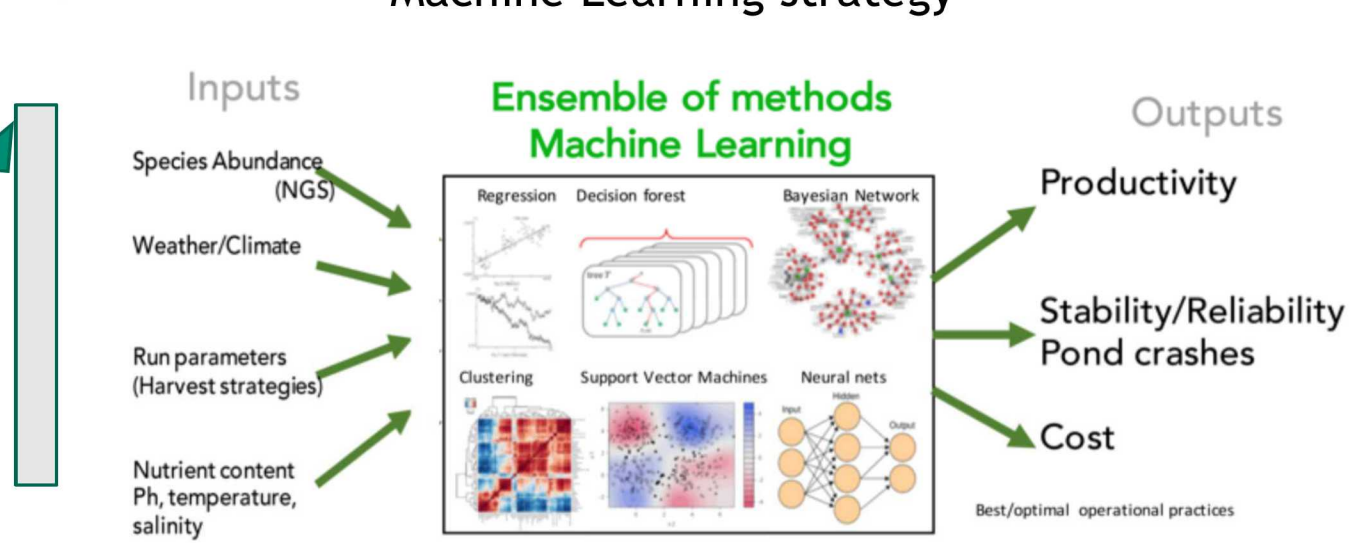
50% more accurate and 10X faster

Classification accuracy

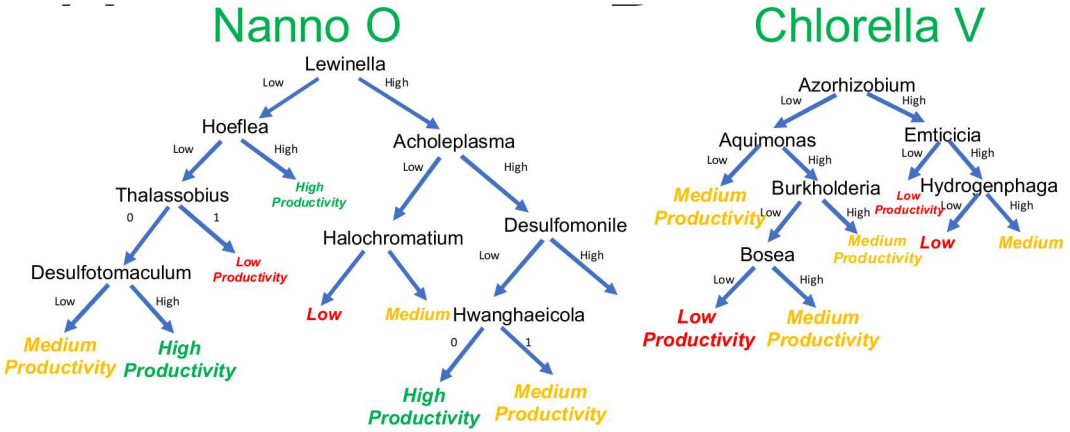
MAGPie		Illumina	
100%	→ Genus	65%	→ Genus
70%	→ Species	30%	→ Species

Machine Learning strategy

Metadata – pond operational data



Development of Predictive Modelling → Anomaly Detection Software

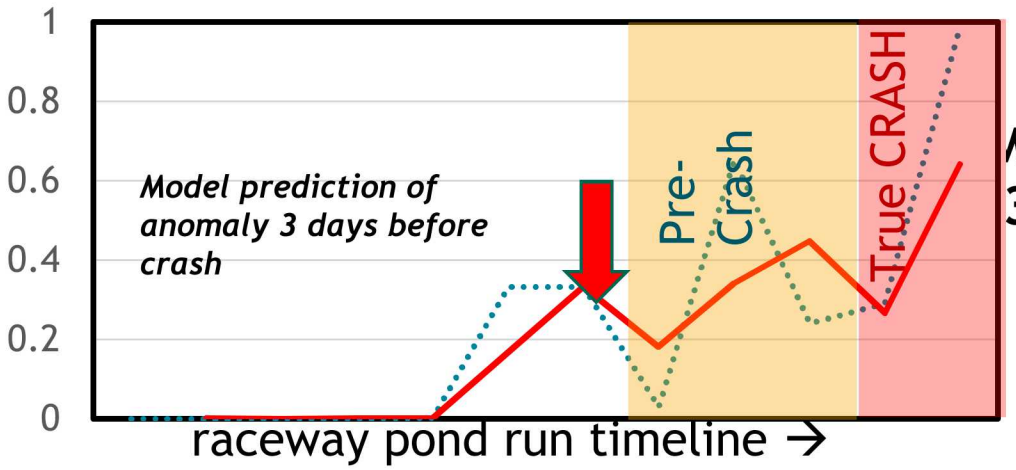


Models	mean accuracy	type
Neural Network - Deep learning	**>97% accuracy**	predictive AFDW
Random Forest	**85% accuracy**	Predictive crash

Feature Importance

Knowledge Based Anomaly Detection

Prediction of anomaly in a cultivation run



Median prediction - 3 days before the crash

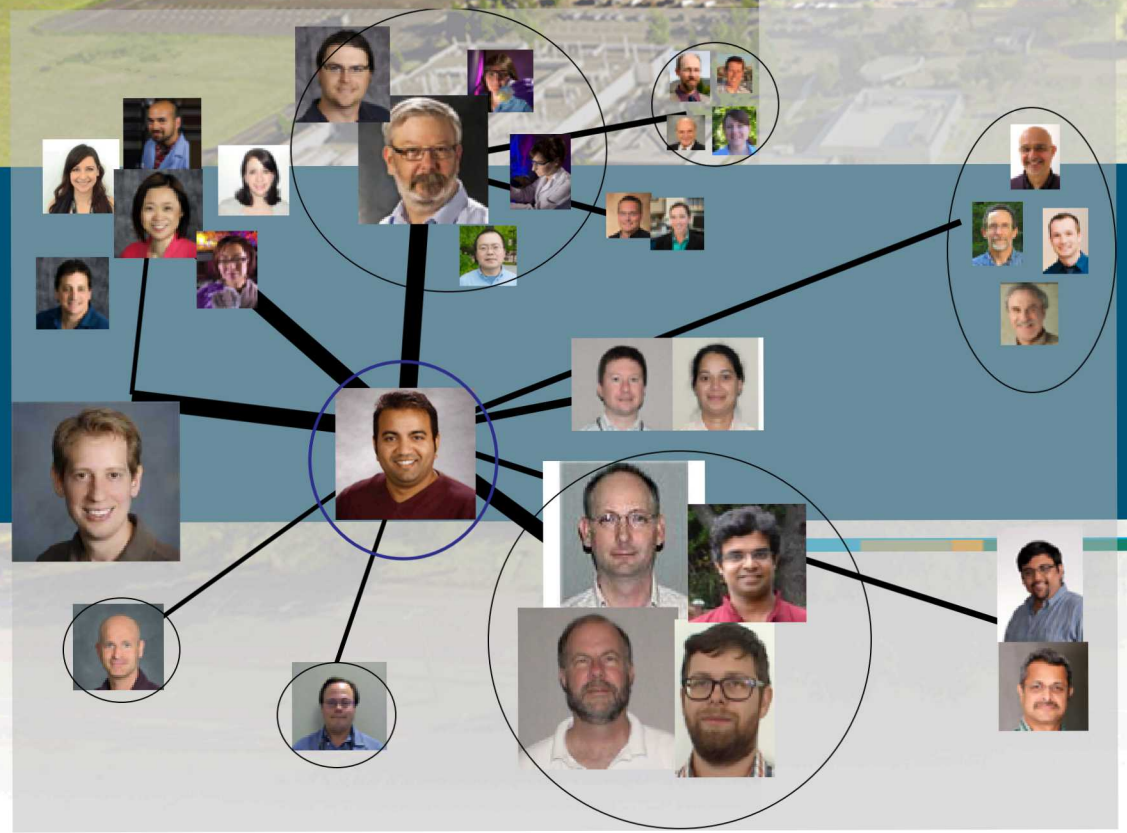
Summary

Here we show a framework of data science tools for analyzing health care related datasets at multiple layers

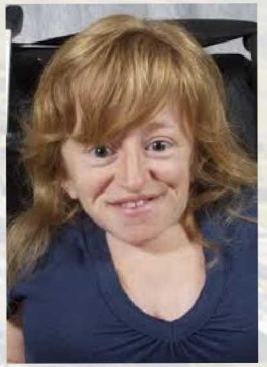
- We showcase our capabilities on successful application of machine learning capabilities on complex biological datasets studying microbiome and multiomics
- Introduce an anomaly detection methods to handle time resolved data streams from multiple sources

And share our vision of how these tools can help in early detection of an emerging biological treat

Thanks for your attention



Acknowledgements



Leanne Whitmore & Prashant M. Rai

Robert Meagher & Ken Sale

Anup Singh