AI EXPERT NETWORK

Explainable and Interpretable AI for medical devices certification

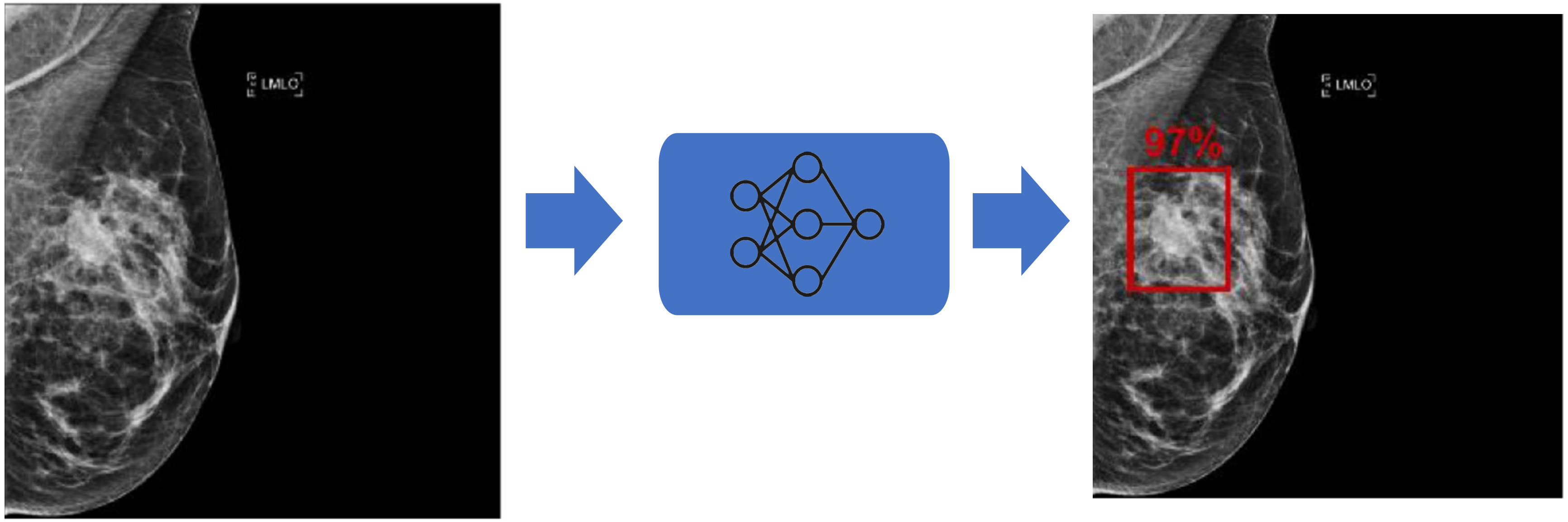
Akhilesh Mishra

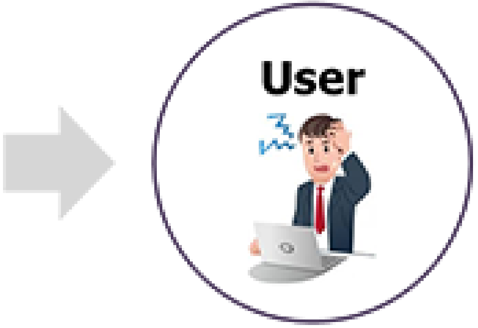
Sr. Medical Devices Industry Manager

MathWorks

Explainable AI / Interpretable AI

Predictions aren’t enough





- Why did you do that?
- Why not something else?
- When do you succeed?
- When do you fail?
- When can I trust you?
- How do I correct an error?

XAI Can :

- Provide hints what is ‘of interest’ to the AI
- Provide hints about weak points of the AI

XAI Can’t :

- Explain the AI
- Replace Good Machine Learning Practices (GMLP), which are key to assuring safety and efficacy

What Clinical Practice expects from AI-powered tools

Improvement, compared to standard practice

- Validate high performance
- Resolve disagreement between AI and human expert (detect systematic error, bias)

Validate prediction against medical knowledge

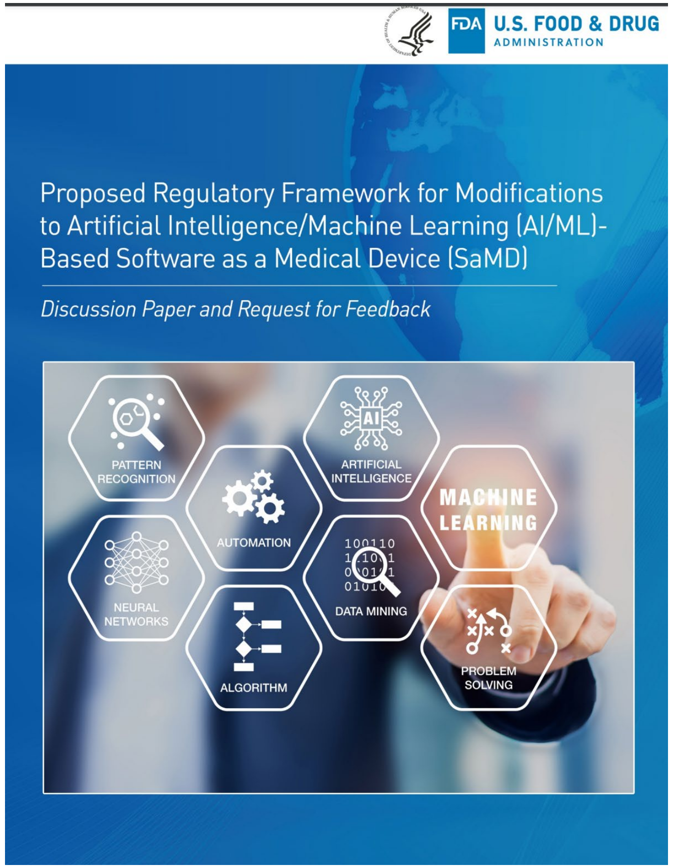
Also explain to patient how recommendations were derived (Patient-centered care)

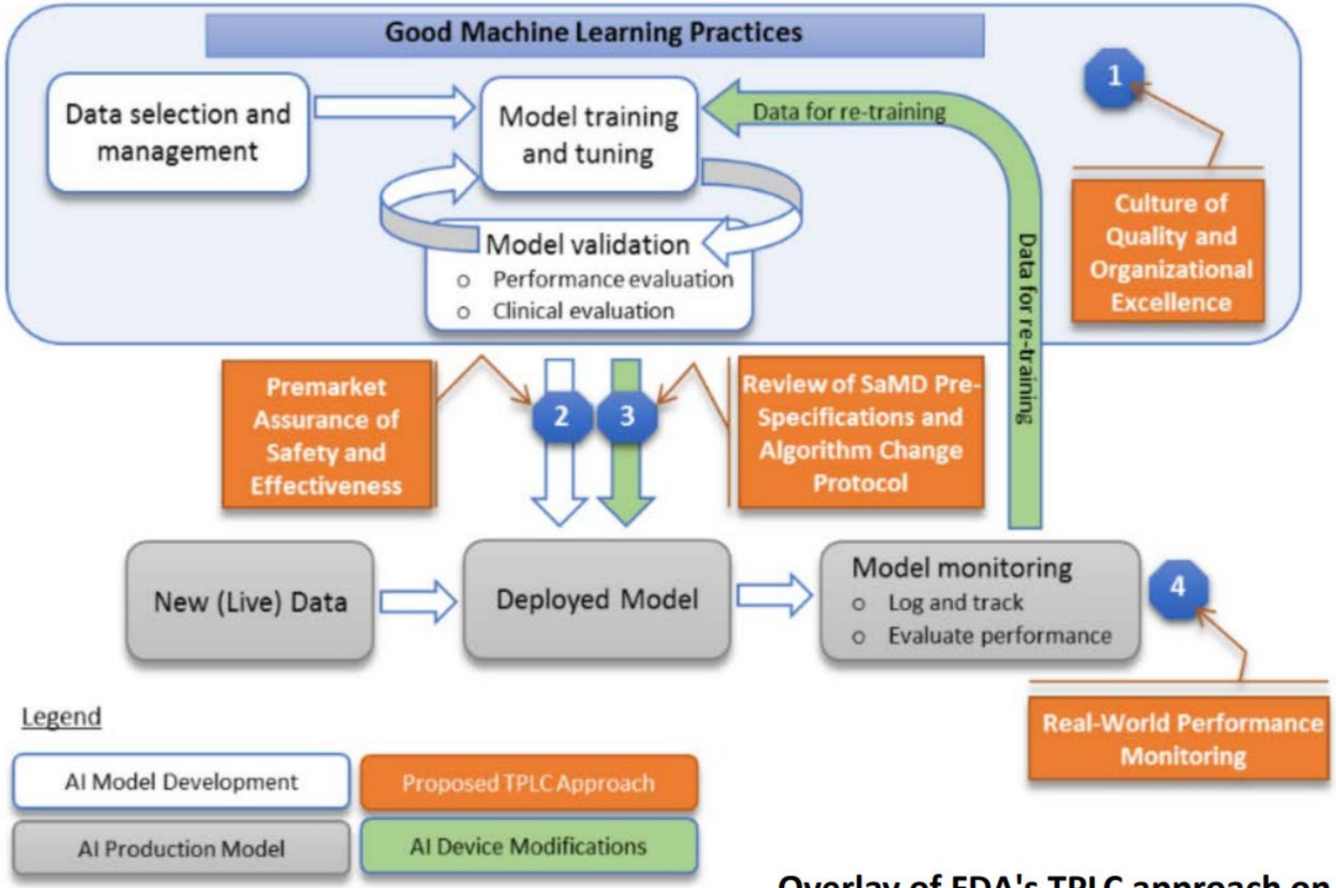
Bad outcomes adequately **mitigated**

- Legal: Empower patient to give informed consent – protect from liability
- Ethical: No bias – Meet beneficence standard (promote optimal outcome for individual)

- *https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7706019/*
- *Good Machine Learning Practice - https://www.fda.gov/media/153486/download*

Where does Explainable AI fit in ?





Overlay of FDA's TPLC approach on AI/ML workflow

https://www.fda.gov/media/122535/download

Key Takeaways: why Explainable AI

- Choosing a method for interpretability based on type of data
- Applying interpretability methods to explain model predictions
- Explainable AI for deep learning and machine learning algorithms
- Certification workflows for AI/ML Software as a Medical Device (SaMD)

Poll

What is your need for explainable AI models?

- Understanding how the model works
- Building trust with stakeholders (clinicians, collaborators, etc.)
- Explaining the model for regulators

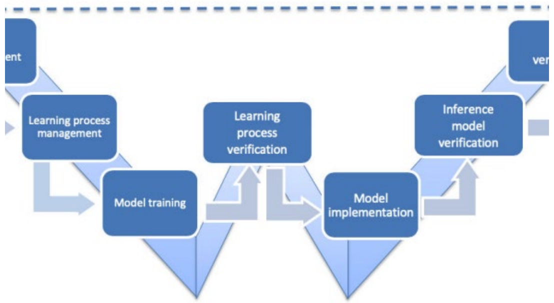
Other

____

We are going to cover the following AI topics:

Interpretability & Explainability

Bias detection and mitigation

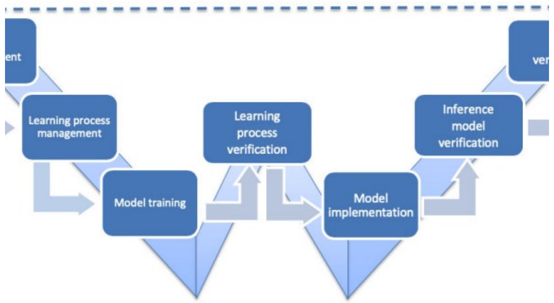


Verification, Validation, & Certification

We are going to cover the following AI topics:

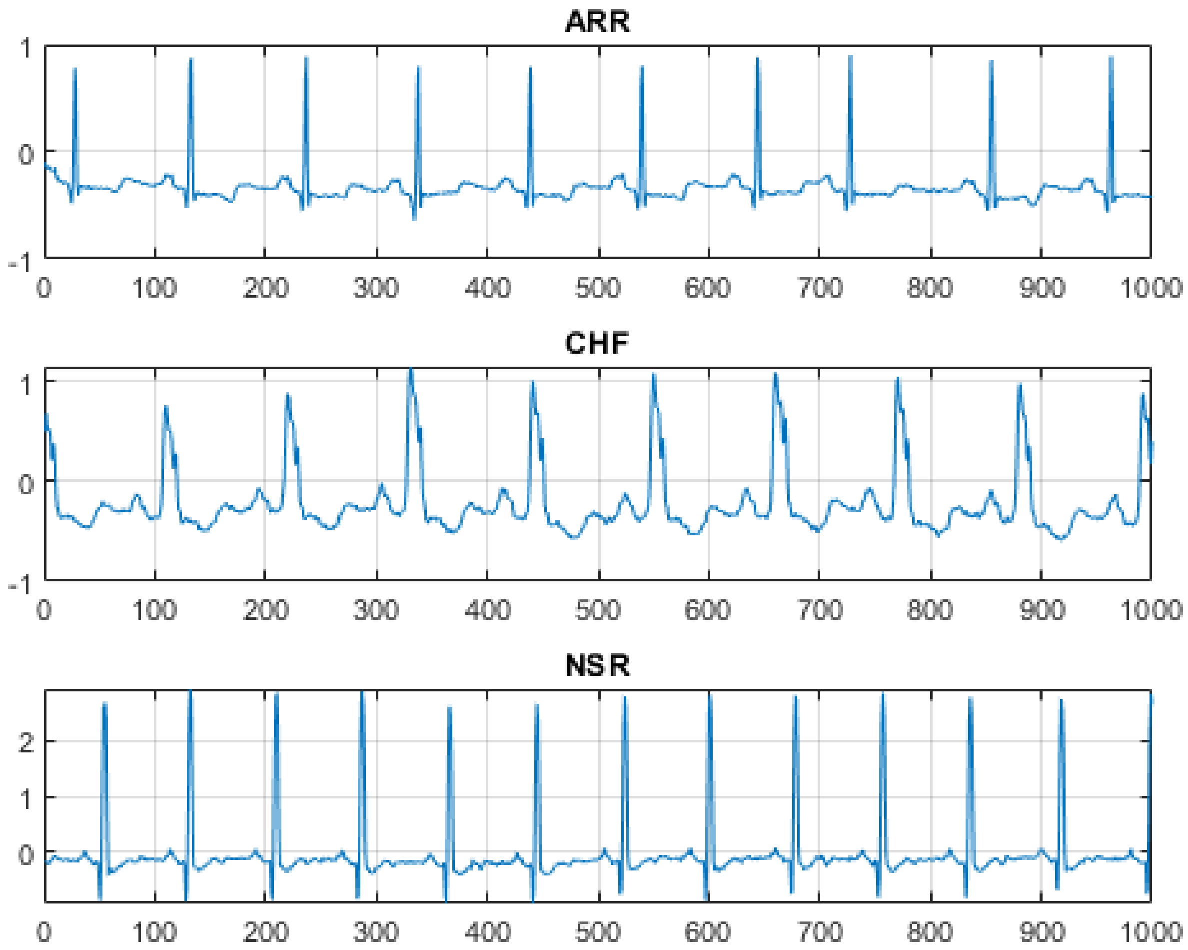
Interpretability & Explainability

Bias detection and mitigation



Verification, Validation, & Certification

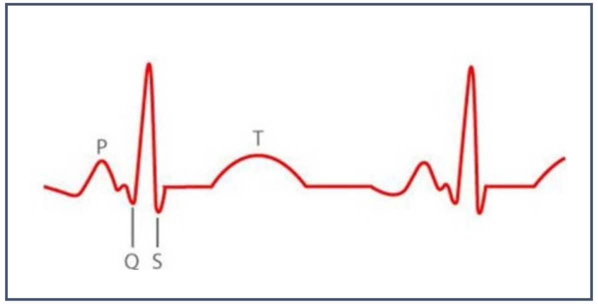
Identifying arrhythmia in ECG data



Classify heartbeat into Normal or Abnormal using ECG recordings in 3 diagnostic categories

https://physionet.org/

Diagnose arrhythmia through Machine Learning



Normal

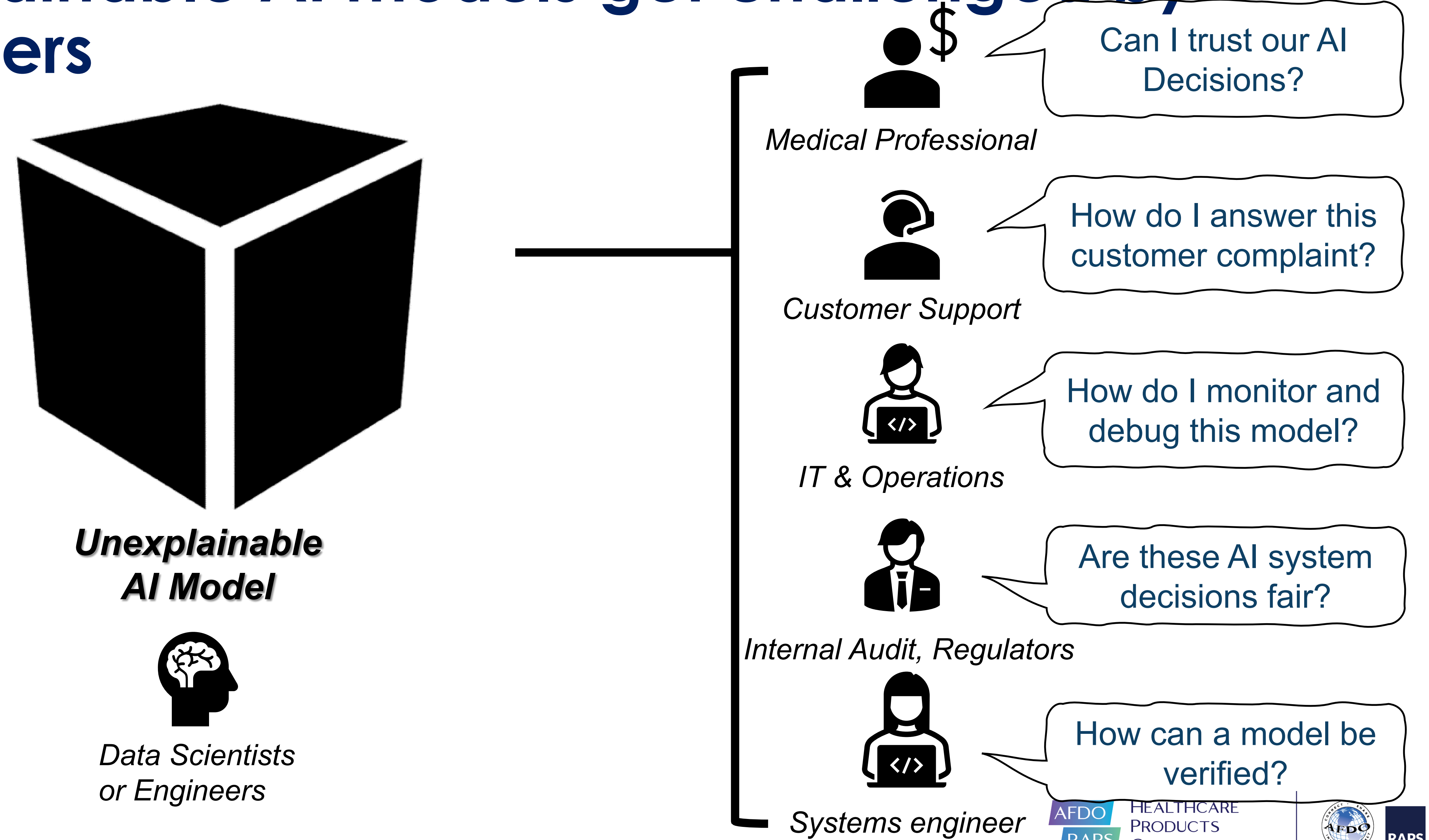
Abnormal

Clean and Preprocess Data → **Engineer Features** → **Select best model and tune hyperparameters** → **Train final model** → **Great performance on test set**

Job Done?

AI EXPERT NETWORK

Unexplainable AI models get challenged by their users

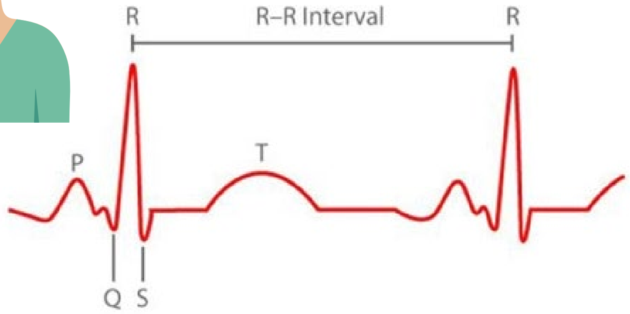


Several methods for Interpretability and Explainability

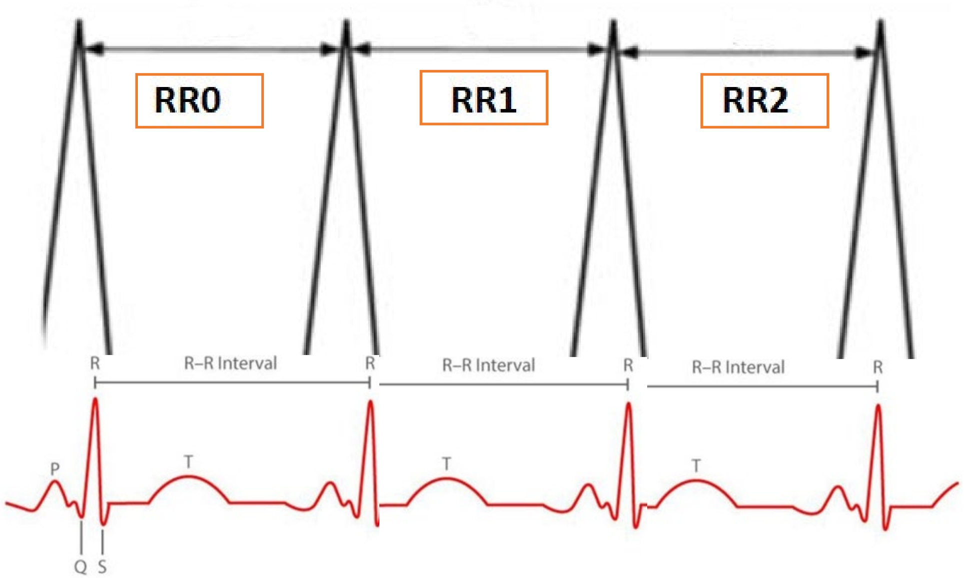
- Interpretability and Explainability Methods
  - Interpretable Models
    - Generalized Linear Models
    - Decision Tree
    - Generalized Additive Models
  - Post Hoc Techniques
    - Tabular Data
      - Partial Dependence Plots
      - Individual Conditional Expectation
      - Shapley
      - LIME
    - Time Series Data
    - Image Data
      - GradCAM
      - Image LIME
      - Deep Dream Image
      - Occlusion Sensitivity

Train a machine learning model for ECG classification

Goal: Classify as Normal or Abnormal



Extract 6 features relating to the R wave



$$R1 = \frac{RR0}{RR1}$$

$$R2 = \frac{RR2}{RR1}$$

$$Rm = \frac{RRmean}{RR1}$$

Fit a Random Forest model:

Accuracy: 99.9%

AI EXPERT NETWORK

Interpret predictions of a machine learning model at a specific point

True Class: Abnormal

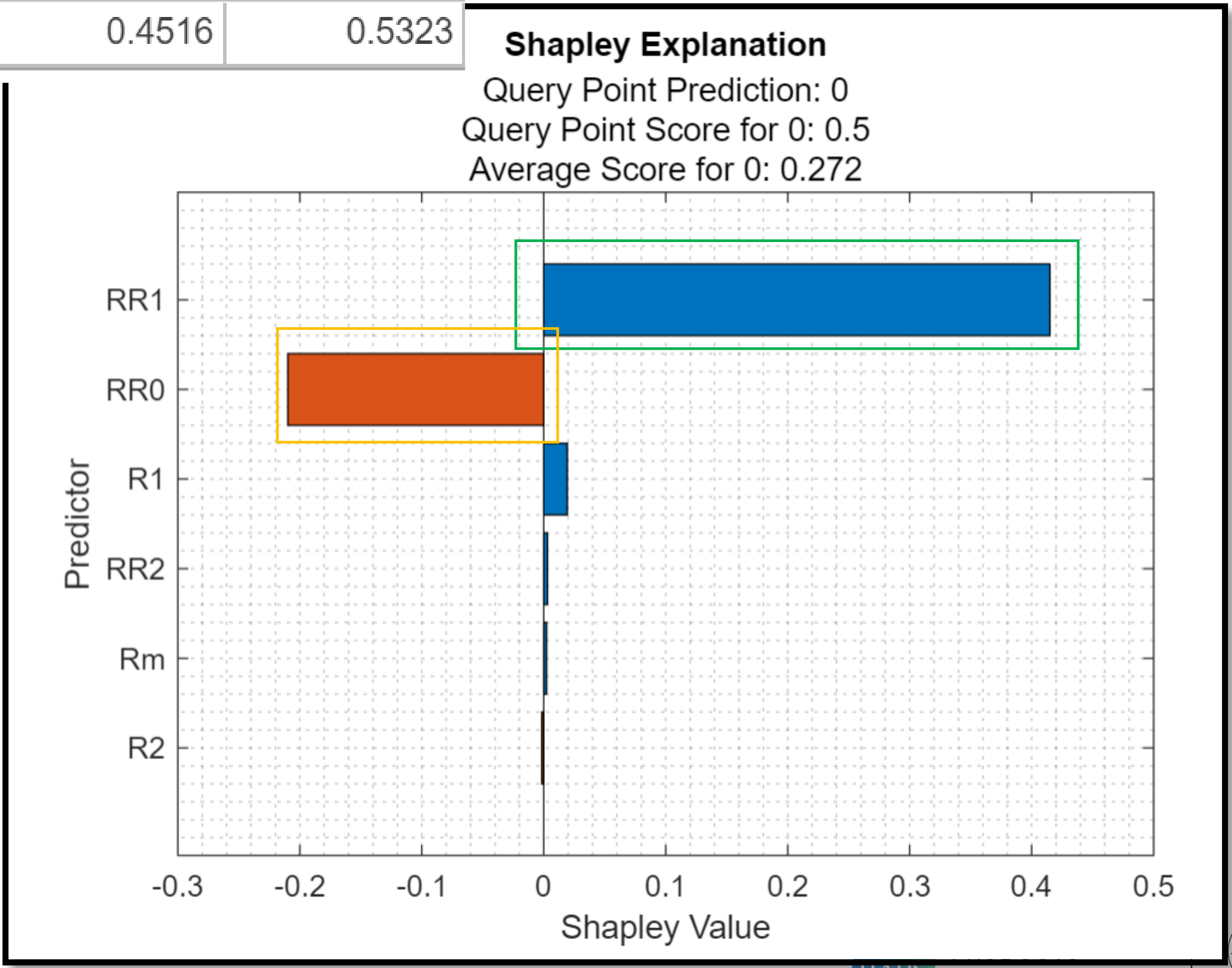
Predicted Class: Normal

	RR0	RR1	RR2	R1	R2	Rm
1	0.0250	0.1722	0.0778	0.1452	0.4516	0.5323

Abnormal Beat (RR0)

Normal Beat (RR1)

Shapley helps us examine individual predictions to see what factors are most important for a specific prediction.



AI EXPERT NETWORK

Poll

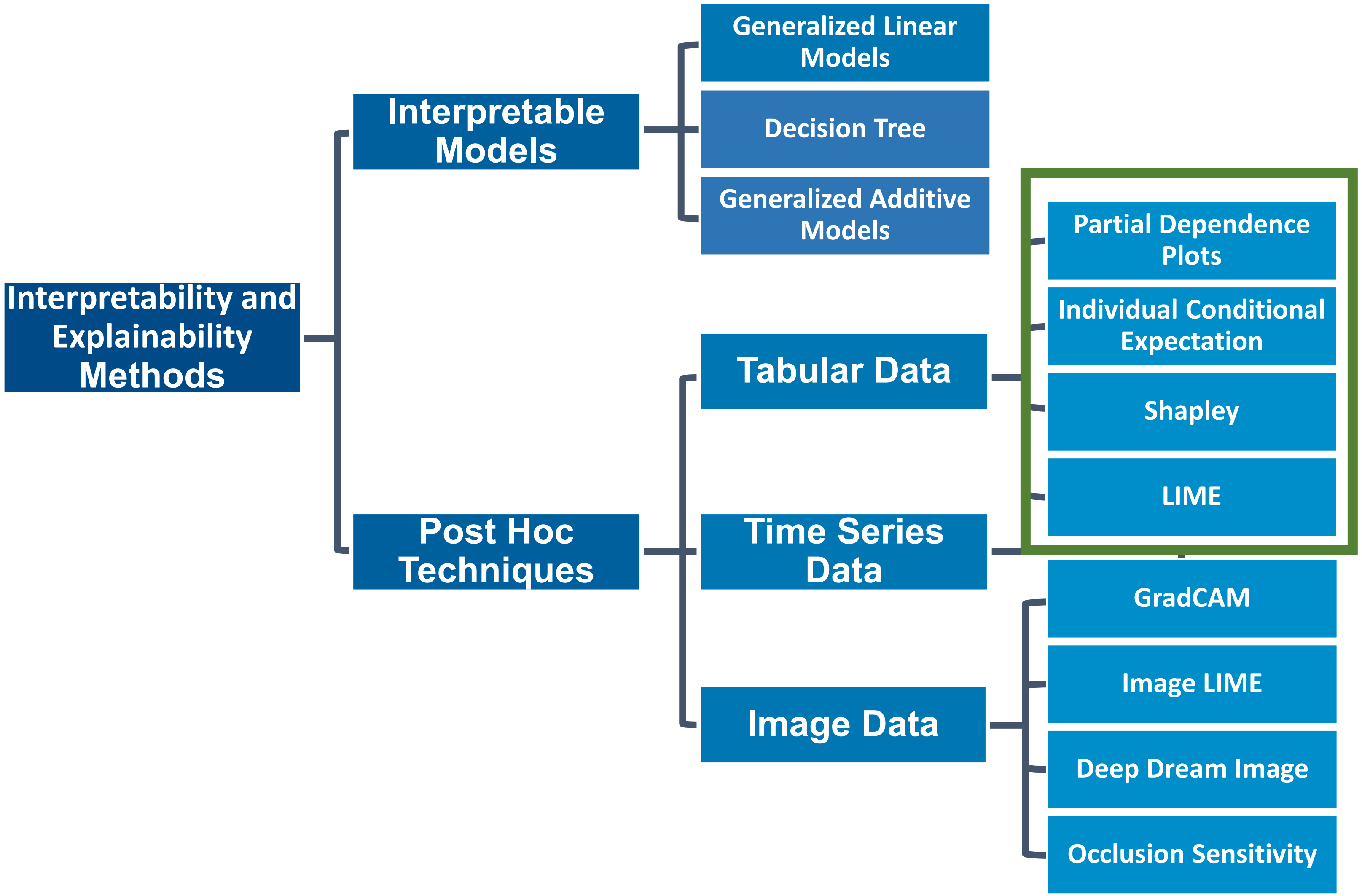
What type of data do you use most?

- Time-series or sequential data
- Tabular data
- Image and video data
- Text data

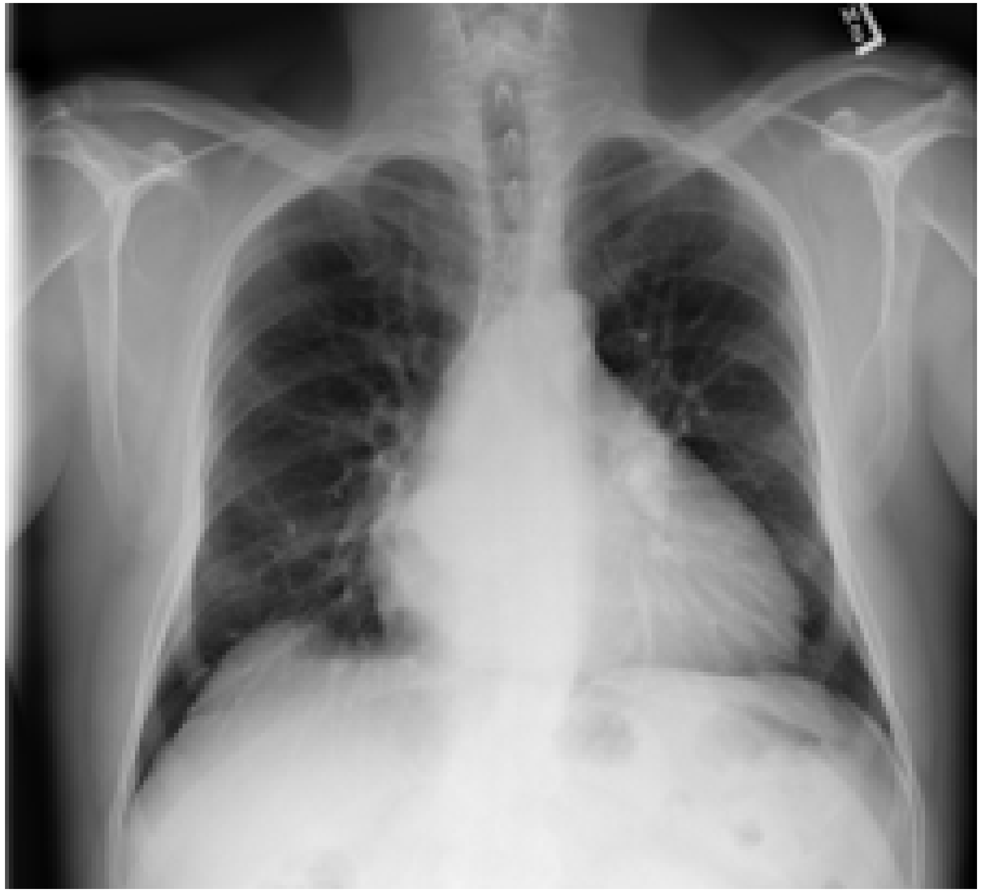
Other

____

Several methods for Interpretability and Explainability



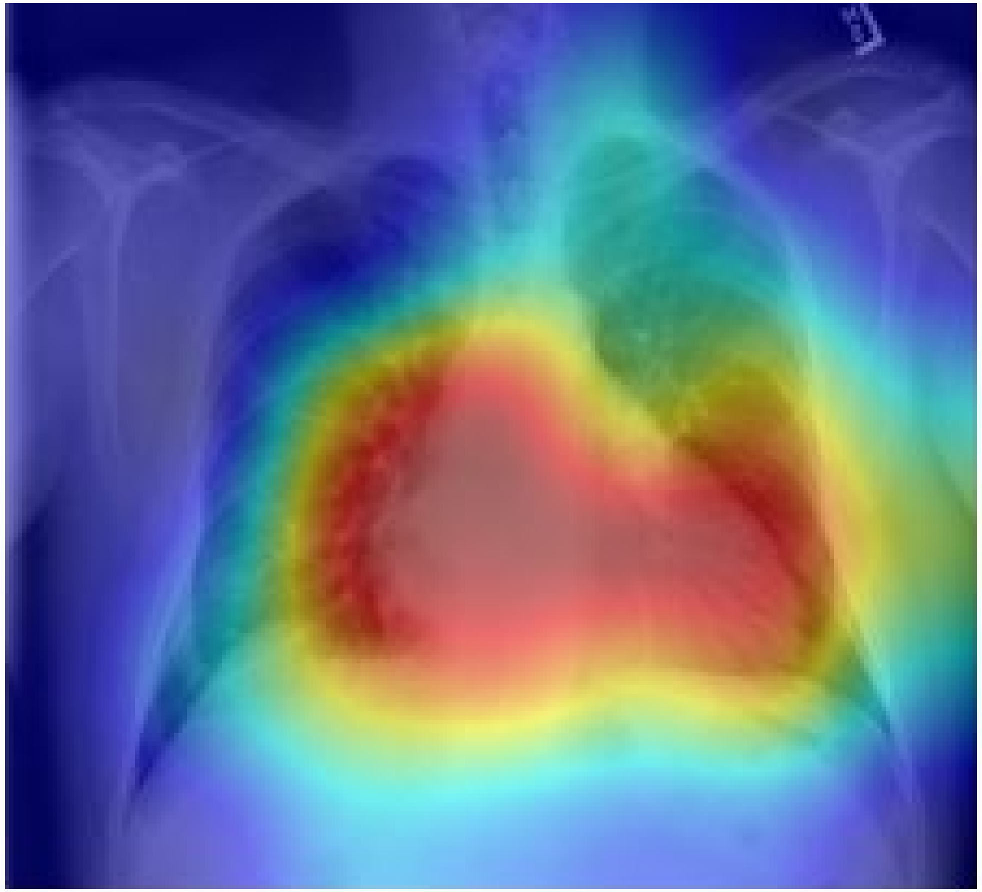
Identifying pathologies in chest Xray images



Identify 14 pathologies from chest Xray's

https://nihcc.app.box.com/v/ChestXray-NIHCC/folder/36938765345

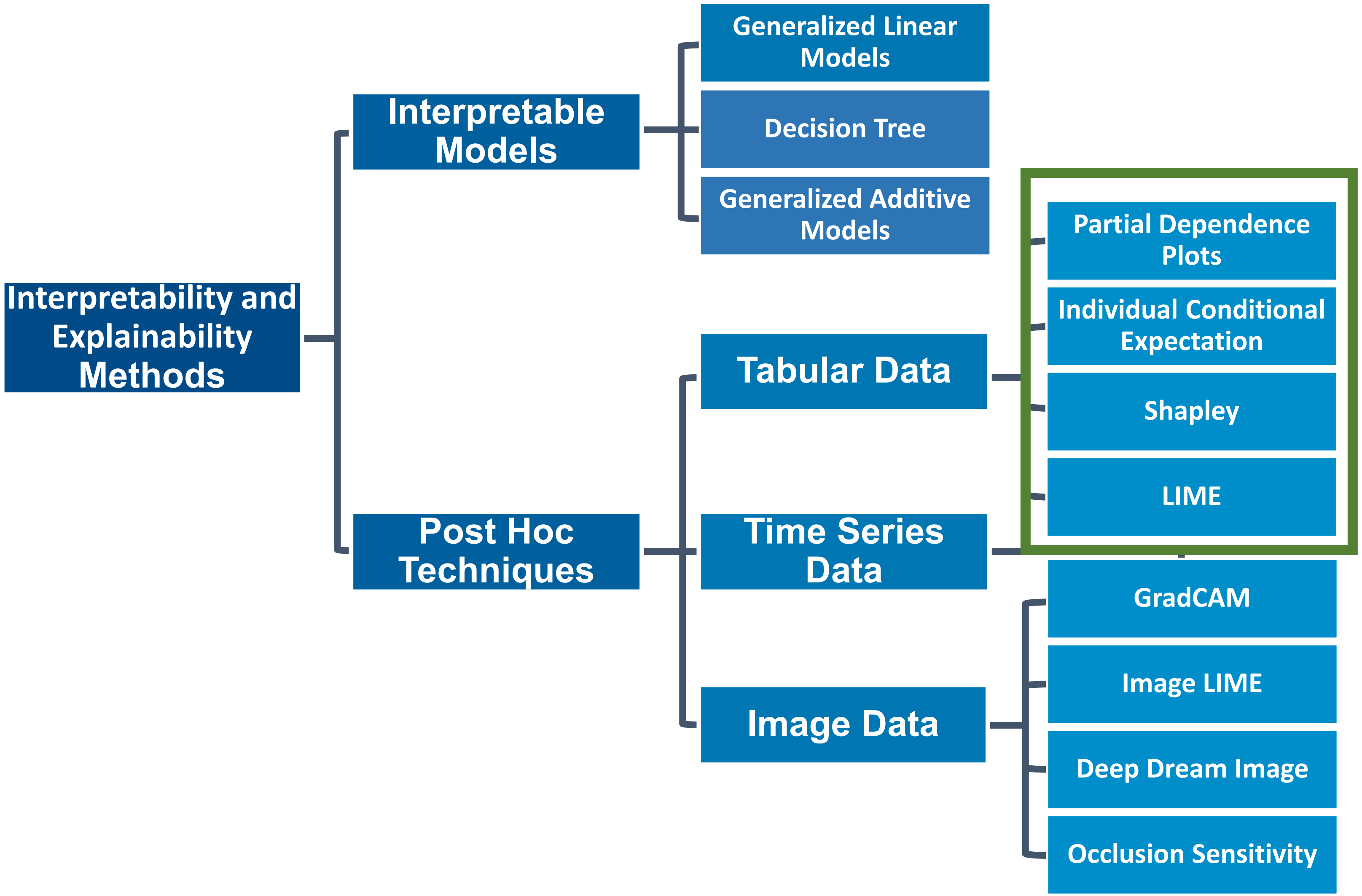
Understanding why the image identifies a pathology



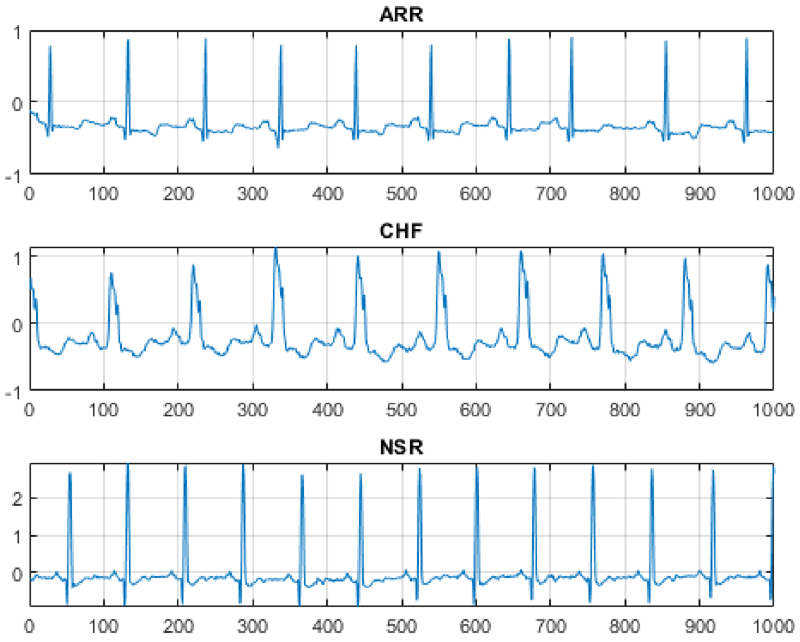
GradCAM shows area of the image that contributes most to classification label

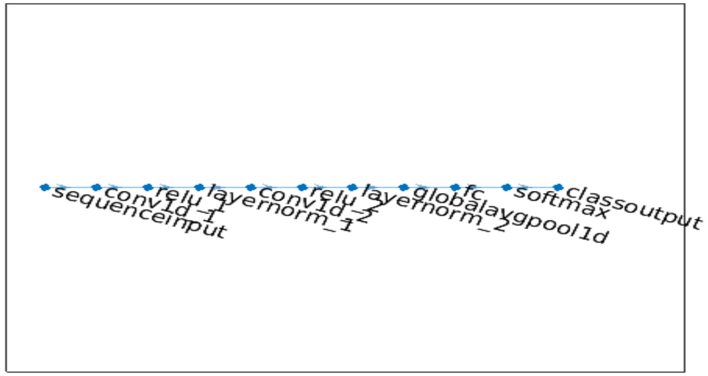
AI Expert Network

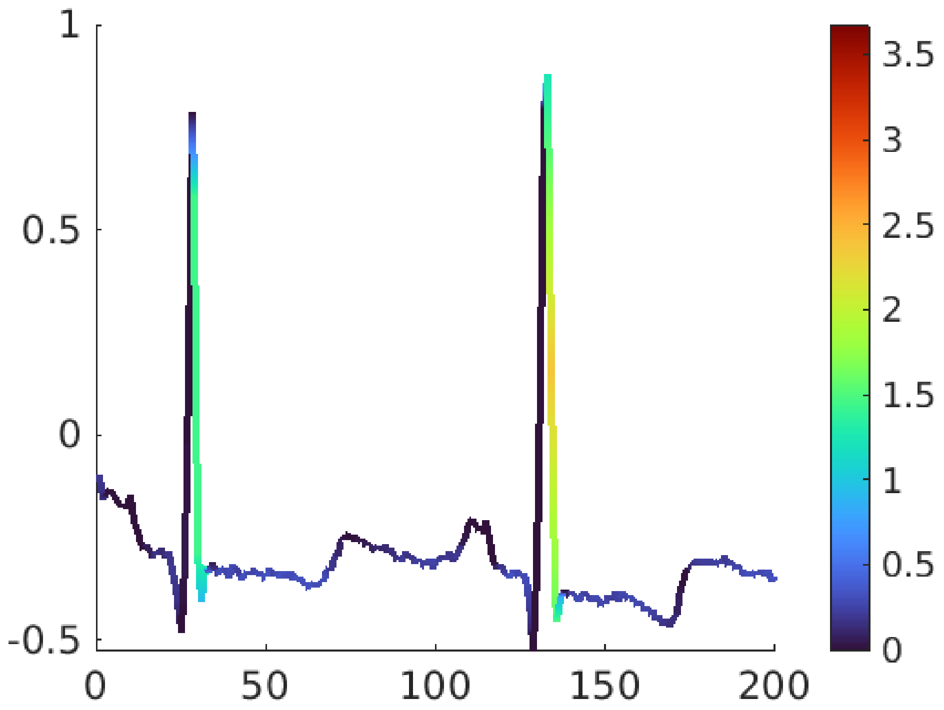
Several methods for Interpretability and Explainability



1D GradCAM reveals why the signal is classified as Arrhythmia



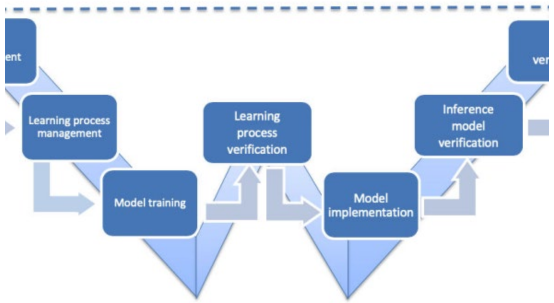




We are going to cover the following AI topics:

Interpretability & Explainability

Bias detection and mitigation



Verification, Validation, & Certification

Bias and Fairness in AI Systems

Source of Bias

Not enough data, Bias through Selection

Legacy bias in the data, Bias through Behavior

Model issues

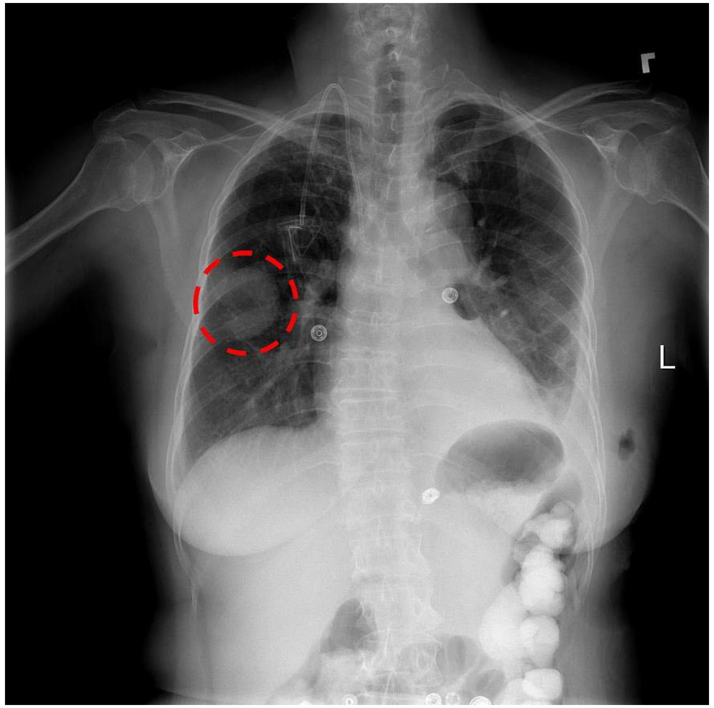
Fairness in Responsible AI: Detecting and mitigating bias against unprivileged groups in ML modeling

AI EXPERT NETWORK

Your devices could be biased!

Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis

Agostina J. Larrazabal[a,1], Nicolás Nieto[a,b,1], Victoria Peterson[b,c], Diego H. Milone[a], and Enzo Ferrante[a,2]



Courtesy : *PNAS*

From oximeters to AI, where bias in medical devices may lurk

Analysis: issues with some gadgets could contribute to poorer outcomes for women and people of colour

Some research suggest that oximeters work less well for patients with darker skin. Photograph: Grace Cary/Getty Images

Courtesy : *The Guardian*

Fixing Medical Devices That Are Biased against Race or Gender

Designers should show how well instruments perform across different populations

Courtesy : *Scientific American*

Problem Formulation

With this model, the intent is to predict a given X-ray image as either normal(no disease associated) or abnormal (have one or more of diseases). This model is thus capable of performing binary classification.

Dataset

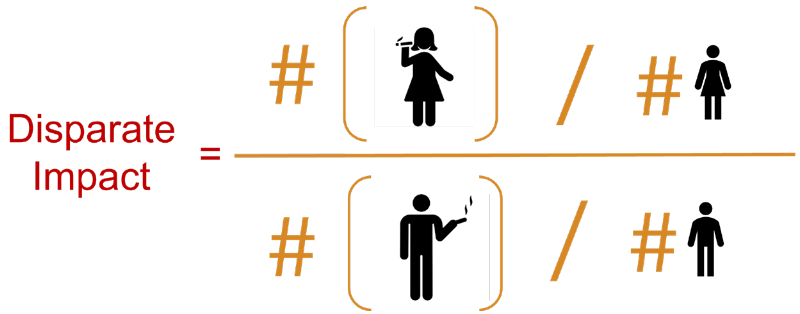
NIH Chest X-ray Dataset is comprised of 112,120 X-ray images with 14 text-mined disease labels from 30,805 unique patients. The 14 diseases labels are *Atelectasis*, *Cardiomegaly*, *Consolidation*, *Edema*, *Effusion*, *Emphysema*, *Fibrosis*, *Hernia*, *Infiltration*, *Mass*, *Nodule*, *Pleural Thickening*, *Pneumonia*, *Pneumothorax*.

We'll take a total of 4961 images

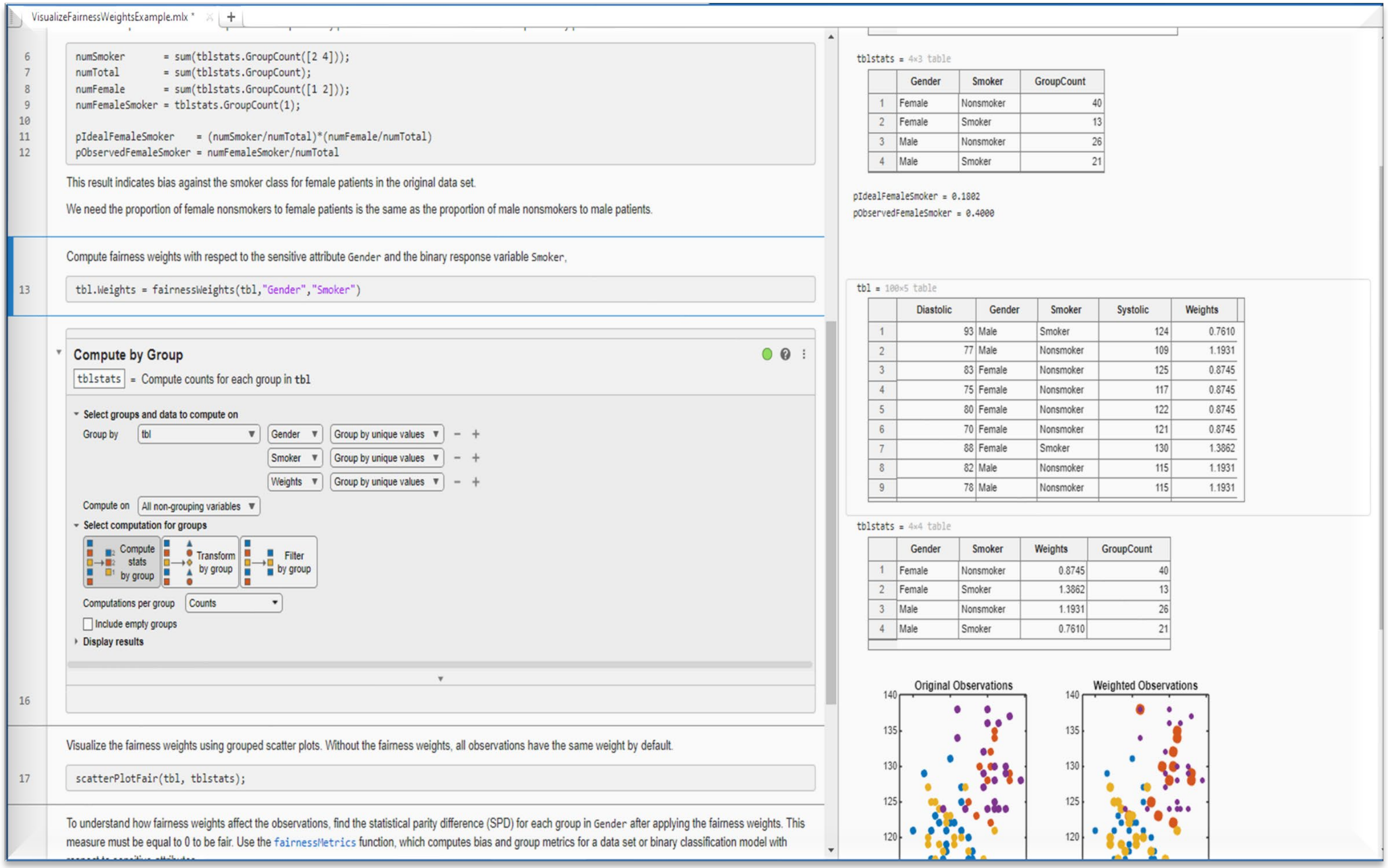
```
tbl            = readtable('Data_Entry_2017_v2020.csv','Range', '1:100000')
```

```
hasCardiomegaly = contains(tbl.FindingLabels, "Cardiomegaly");
sum(hasCardiomegaly)

noFinding = contains(tbl.FindingLabels,"No Finding");
sum(noFinding)

isHealthy        = false(numel(noFinding), 1);
```

Measure fairness - Detect bias



Disparate Impact < 1 for females indicating bias



	Gender	Smoker	GroupCount
1	Female	Nonsmoker	40
2	Female	Smoker	13
3	Male	Nonsmoker	26
4	Male	Smoker	21

	Diastolic	Gender	Smoker	Systolic	Weights
1	93	Male	Smoker	124	0.7610
2	77	Male	Nonsmoker	109	1.1931
3	83	Female	Nonsmoker	125	0.8745
4	75	Female	Nonsmoker	117	0.8745
5	80	Female	Nonsmoker	122	0.8745
6	70	Female	Nonsmoker	121	0.8745
7	88	Female	Smoker	130	1.3862
8	82	Male	Nonsmoker	115	1.1931
9	78	Male	Nonsmoker	115	1.1931

	Gender	Smoker	Weights	GroupCount
1	Female	Nonsmoker	0.8745	40
2	Female	Smoker	1.3862	13
3	Male	Nonsmoker	1.1931	26
4	Male	Smoker	0.7610	21

AI EXPERT NETWORK

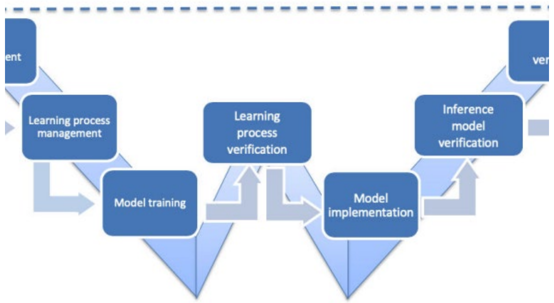
Bias detection and mitigation

Stage	Description
Pre-processing	Removes the information correlated to the sensitive attribute
In-processing	Add constraint or regularization term to the objective, Adversarial models
Post-Processing	Edit posteriors to satisfy fairness constraints

We are going to cover the following AI topics:

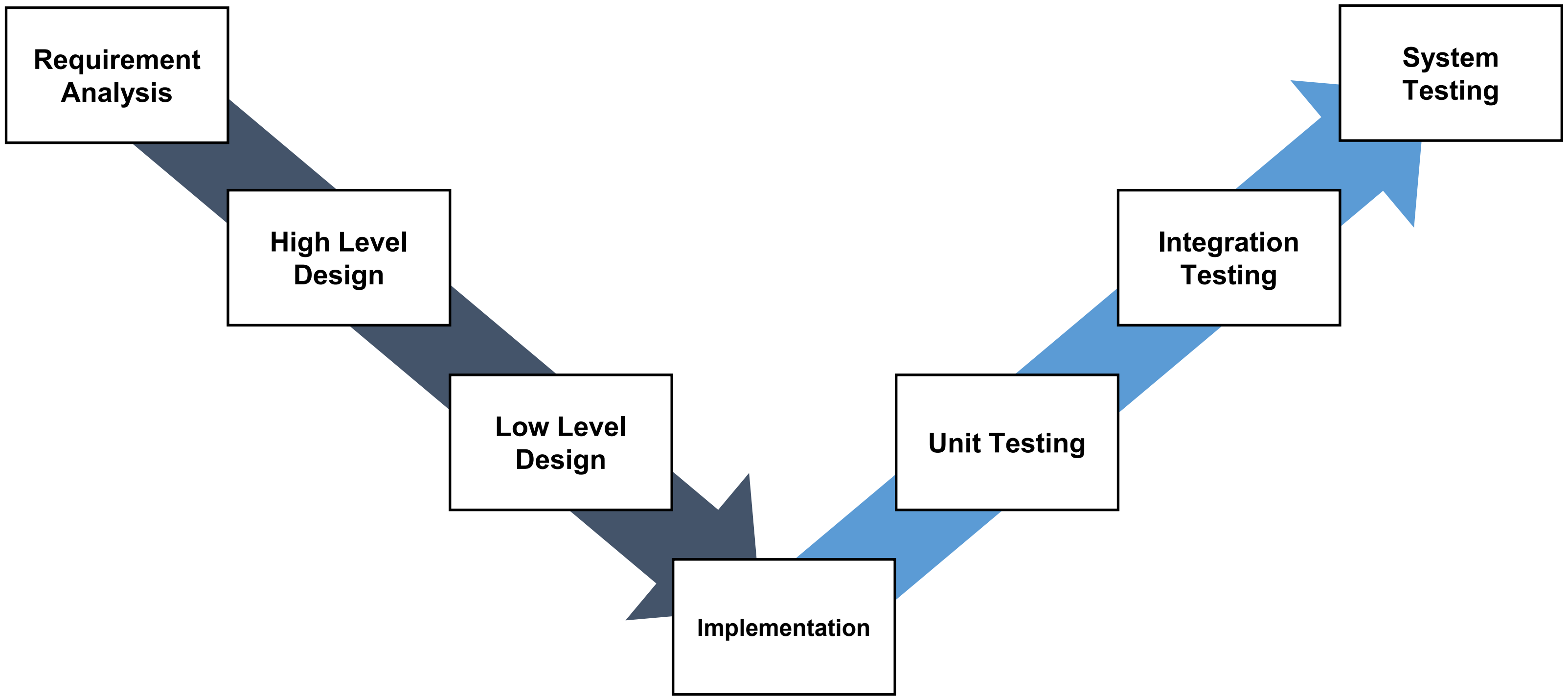
Interpretability & Explainability

Bias detection and mitigation



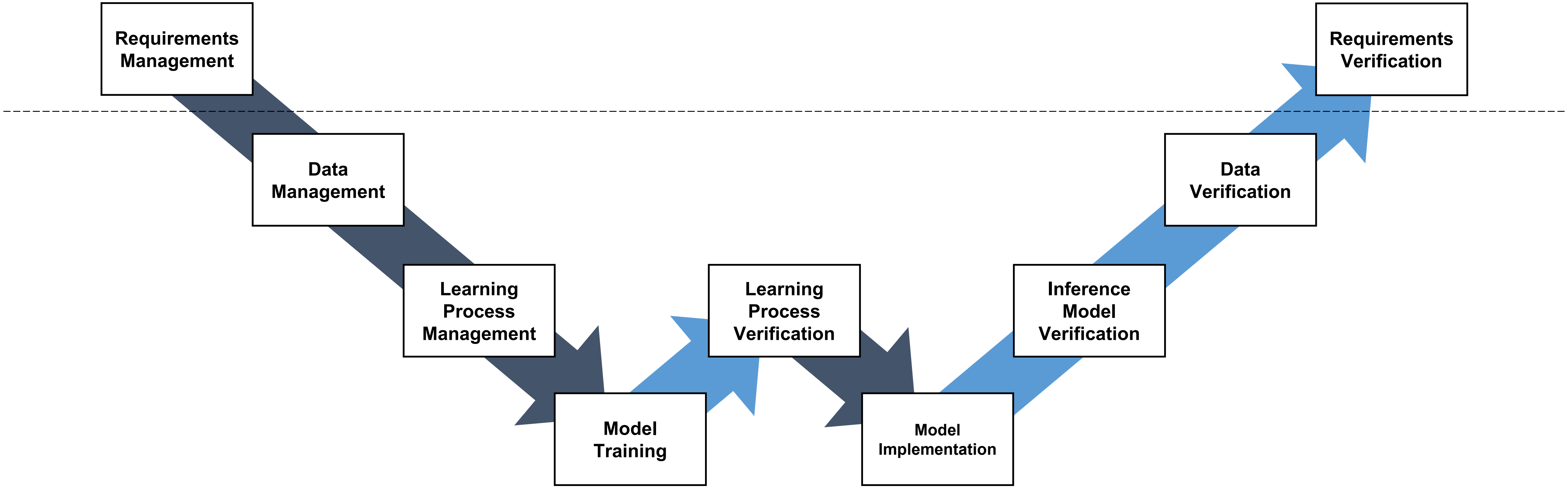
Verification, Validation, & Certification

Traditional medical device development uses verification and validation



Example: An ECG Heart Rate Monitor would follow IEC 62304

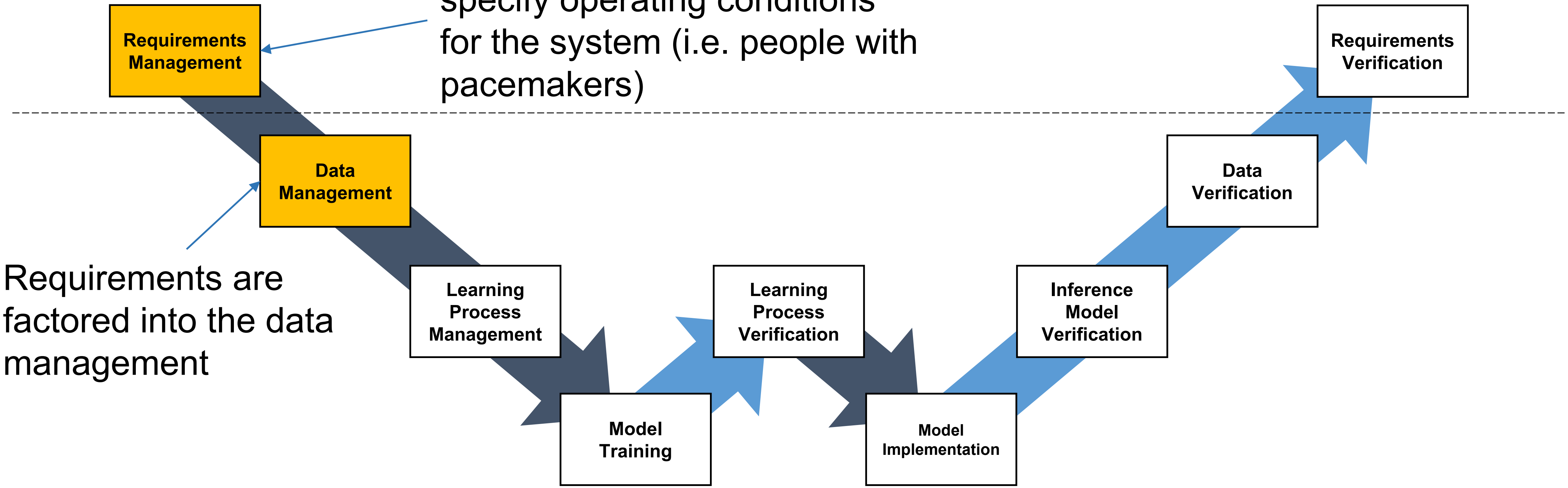
The V-diagram can be adapted into the W-diagram to include AI components



[EASA Concepts of Design Assurance for Neural Networks (CoDANN)]

The W-diagram shows how to factor data into verification and validation

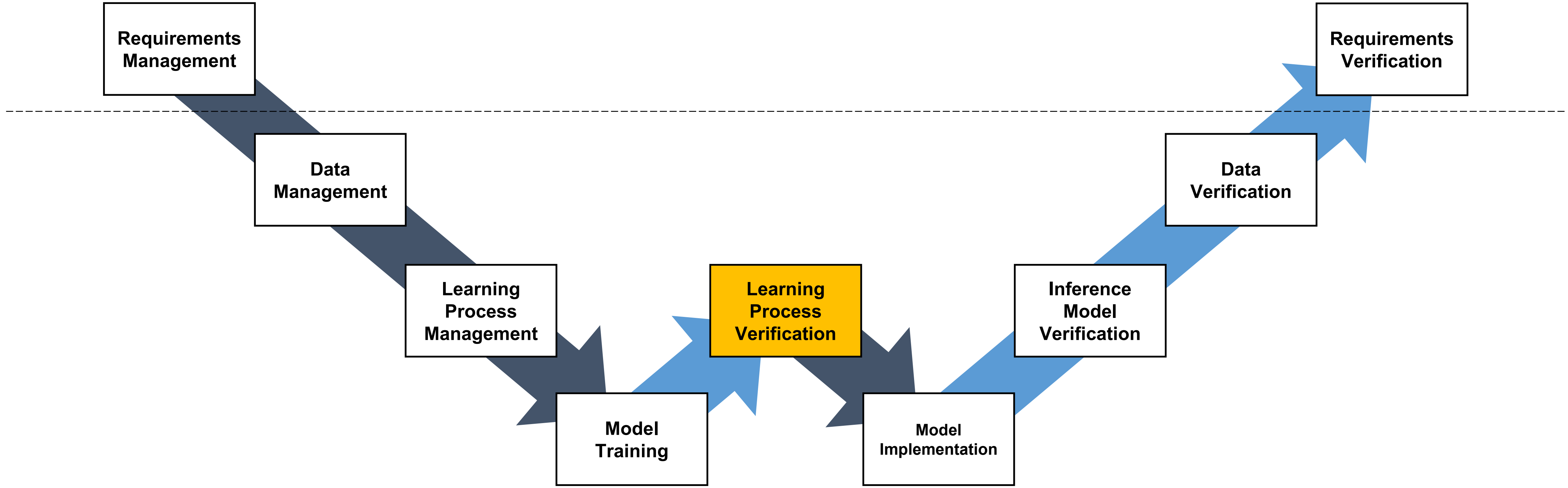
Define requirements that specify operating conditions for the system (i.e. people with pacemakers)



Requirements are factored into the data management

[EASA Concepts of Design Assurance for Neural Networks (CoDANN)]

An important step in the W-diagram is Learning Process Verification



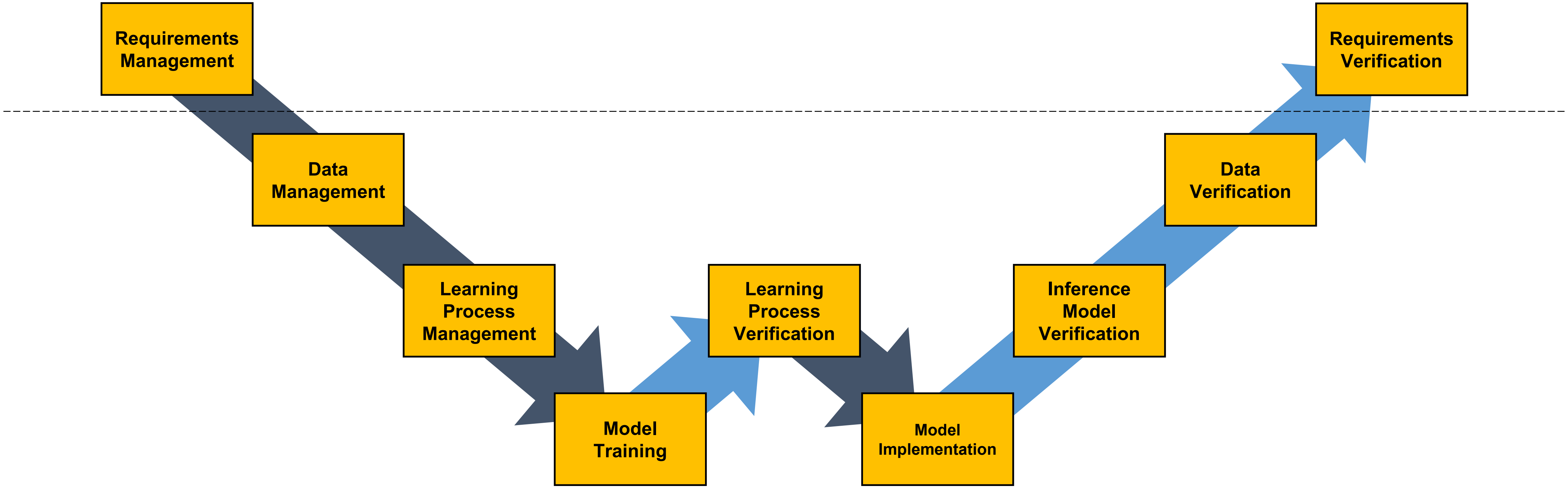
[EASA Concepts of Design Assurance for Neural Networks (CoDANN)]

Poll

Are model explainability/interpretability properties sufficient in your current AI framework?

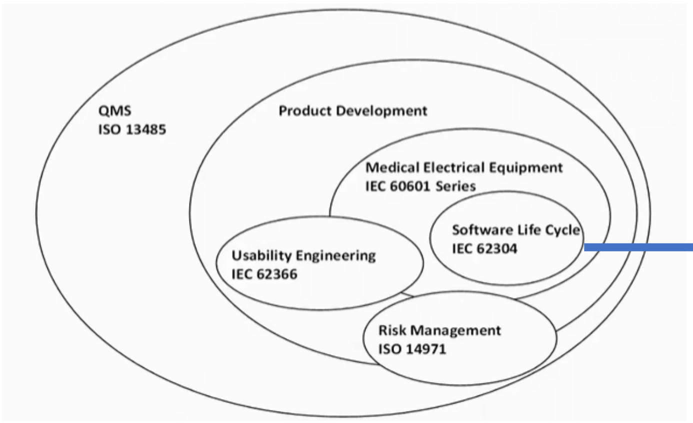
- Yes
- No

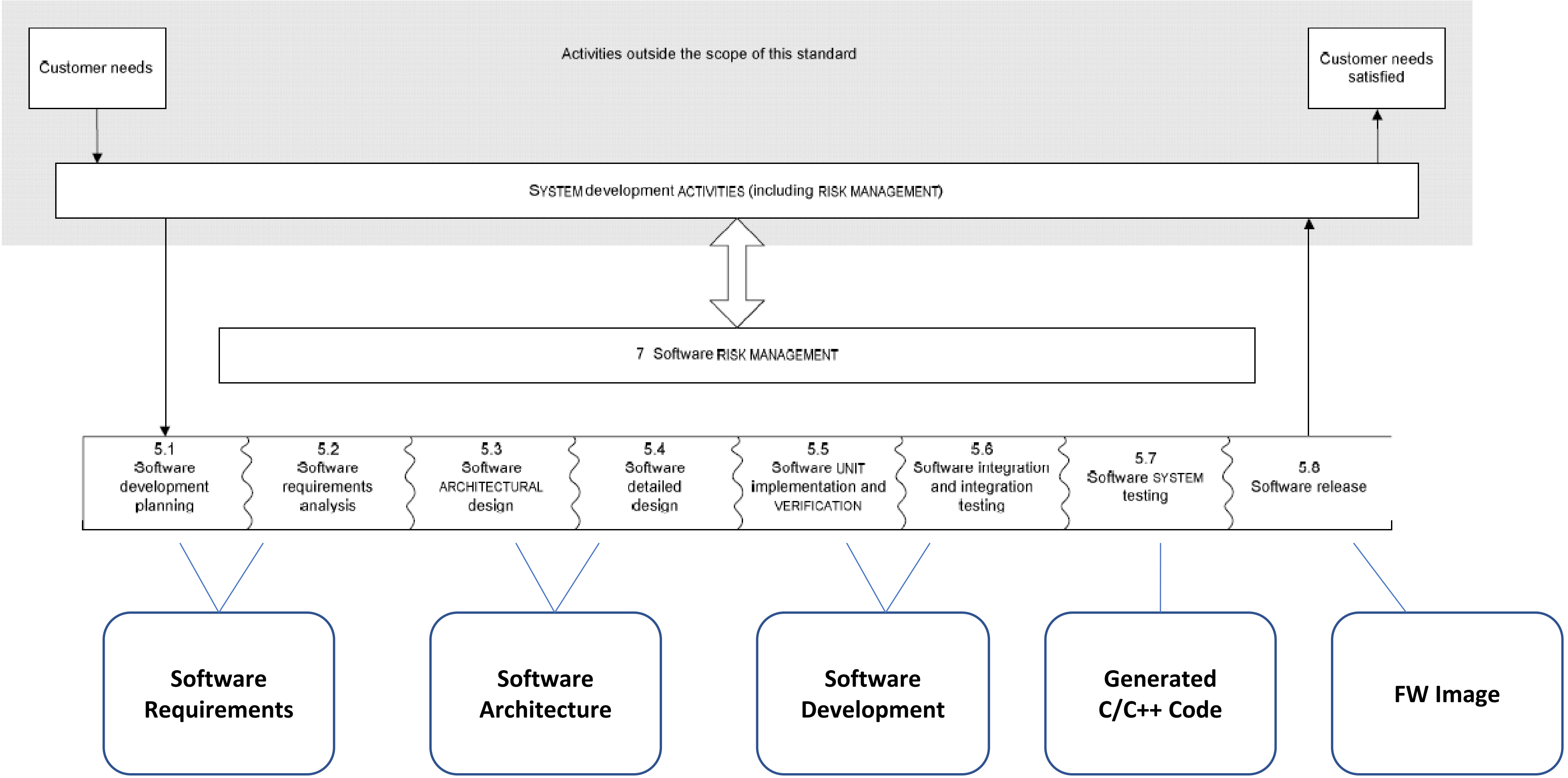
Certification will provide guidance for development of safety critical AI systems



EASA Concepts of Design Assurance for Neural Networks (CoDANN)

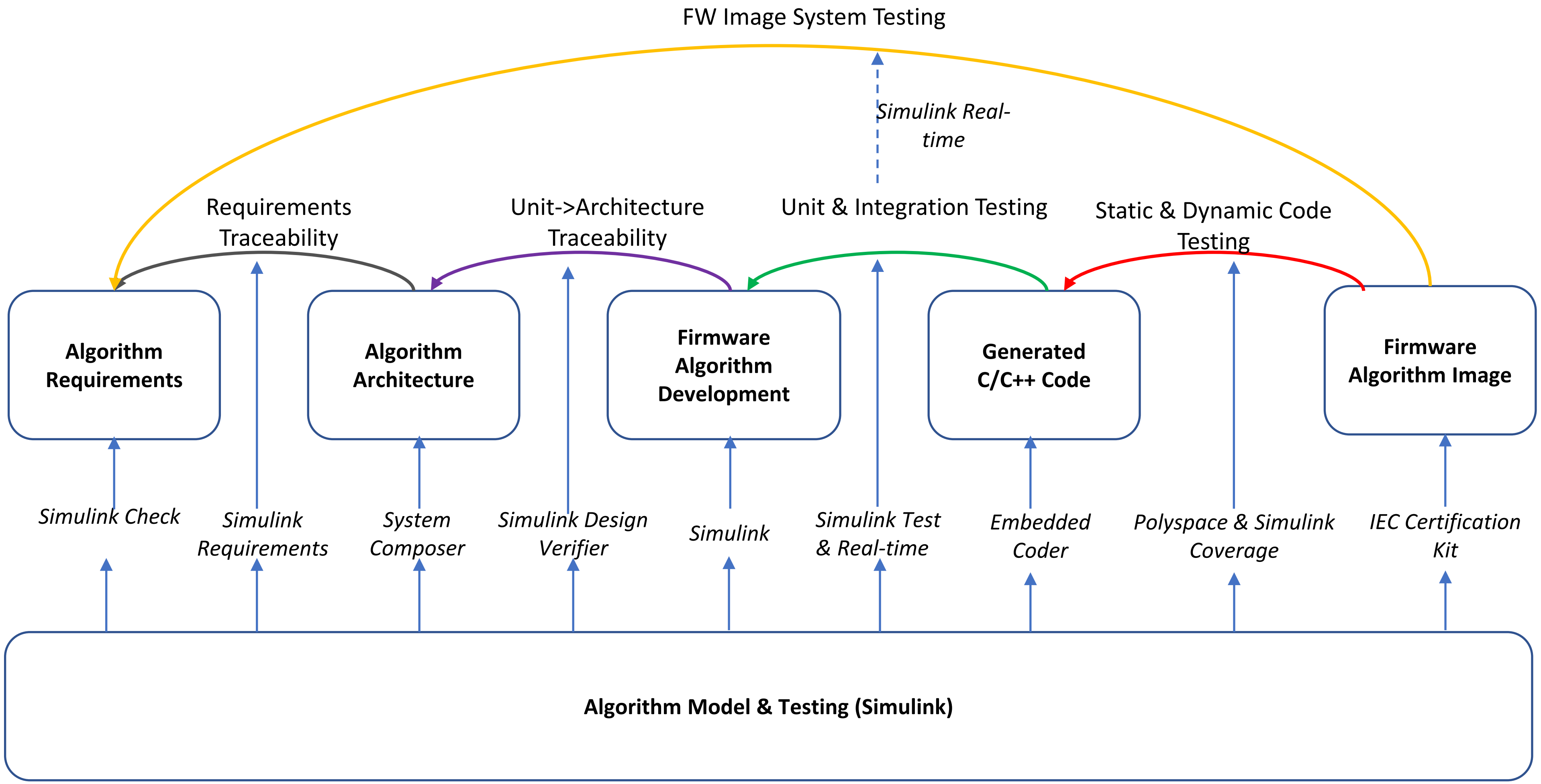
IEC 62304 Breakdown





Section 5 establishes framework for how to develop and test medical software

Algorithm development workflow



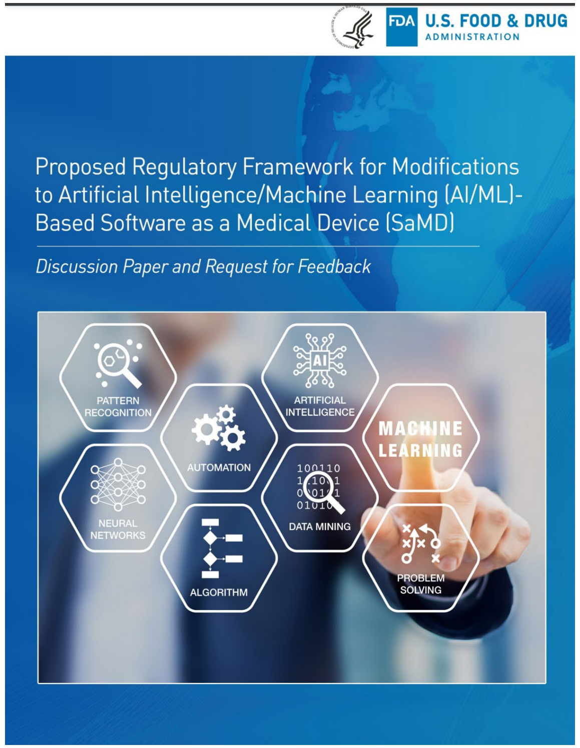
Verification of requirements, architecture, development, and testing are synchronized for both model and generated FW code

Certification for AI in Medical Devices is in the early stages


FDA U.S. FOOD & DRUG ADMINISTRATION
CENTER FOR DEVICES & RADIOLOGICAL HEALTH

Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD) Action Plan

January 2021


FDA U.S. FOOD & DRUG ADMINISTRATION

Proposed Regulatory Framework for Modifications to Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD)

Discussion Paper and Request for Feedback

Technical Performance Assessment of Quantitative Imaging in Radiological Device Premarket Submissions

Guidance for Industry and Food and Drug Administration Staff

JUNE 2022

Download the Final Guidance Document | Read the Federal Register Notice

Final

Share | Tweet | LinkedIn | Email | Print

Docket Number: FDA-2019-D-1470
Issued by: Center for Devices and Radiological Health

This guidance document provides FDA's recommendations on the information, technical performance assessment, and user information that should be included in a premarket submission for radiological devices that include quantitative imaging functions. The recommendations reflect current review practices and are intended to promote consistency and facilitate efficient review of premarket submissions for radiological devices that include quantitative imaging functions.

bsi.

AAMI Advancing Safety in Health Technology

FDA

AI EXPERT NETWORK

Helpful links shared during this session :

- LIME function with examples : https://www.mathworks.com/help/stats/lime.html
- Shapley function with examples: https://www.mathworks.com/help/stats/shapley.html
- GradCAM function with examples : https://www.mathworks.com/help/deeplearning/ref/gradcam.html
- Occlusion sensitivity with examples: https://www.mathworks.com/help/deeplearning/ref/occlusionsensitivity.html
- Verify adversarial robustness of deep learning networks with examples : https://www.mathworks.com/help/deeplearning/deep-learning-verification.html
- Partial dependence plot with examples : https://www.mathworks.com/help/stats/regressiontree.plotpartialdependence.html
- Predictor importance methods for feature selection and explainability : https://www.mathworks.com/help/stats/dimensionality-reduction.html

Thank you !

- Contact email:

Akhilesh Mishra

amishra@mathworks.com

- LinkedIn:

https://www.linkedin.com/in/akhilesh-mishra-mathworks/