

# Human-Centered and Explainable AI: Foundations and Applications in Medicine

**Rayan Ebnali Harari, PhD**

Harvard Medical School | MGB

What is Human-Centered and Explainable AI (xAI) Broadly?

# Designing AI that **makes sense** to people

# Agenda

- Part I: Foundations of Explainable AI (XAI)
- Part II: Explainable Models & Techniques
- Part III: Readmission Risk Use Case
- Part IV: Why XAI Matters in Clinical Practice
- Part V: 7 Aspects of Healthcare XAI
- Part VI: Designing Human-Centered XAI
- Part VII: Case Study & Best Practices

# Gaps in Healthcare



BOSTON GLOBE STAFF ILLUSTRATION/ADOBE

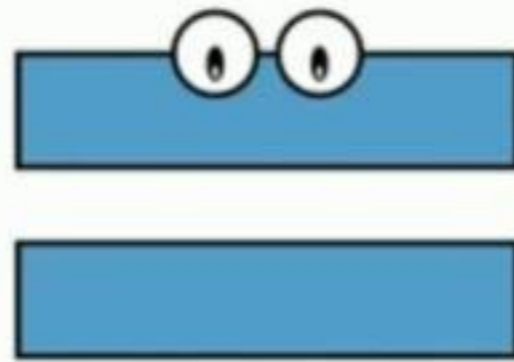


MEDICAL SCHOOL

[maran@bwh.harvard.edu](mailto:maran@bwh.harvard.edu)

*Medical errors are  
responsible for  
between 250,000  
to 400,000  
deaths per year in  
the U.S. alone*

Medical errors are responsible for between 250,000 to 400,000 deaths per year in the U.S. alone



747 plane crashing every single day



# The Cost of Fragmentation in Healthcare

- \$320 B annual waste due to inefficiencies and poor coordination (CMS 2024)
- Clinicians face burnout, cognitive overload
- Critical decisions are delayed or missed in high-risk setting
- Rural and underserved communities suffer disproportionately
- We need context-aware, human-centered medical tech to close these gaps





# Lane 1: AI for Clinical Decision Support

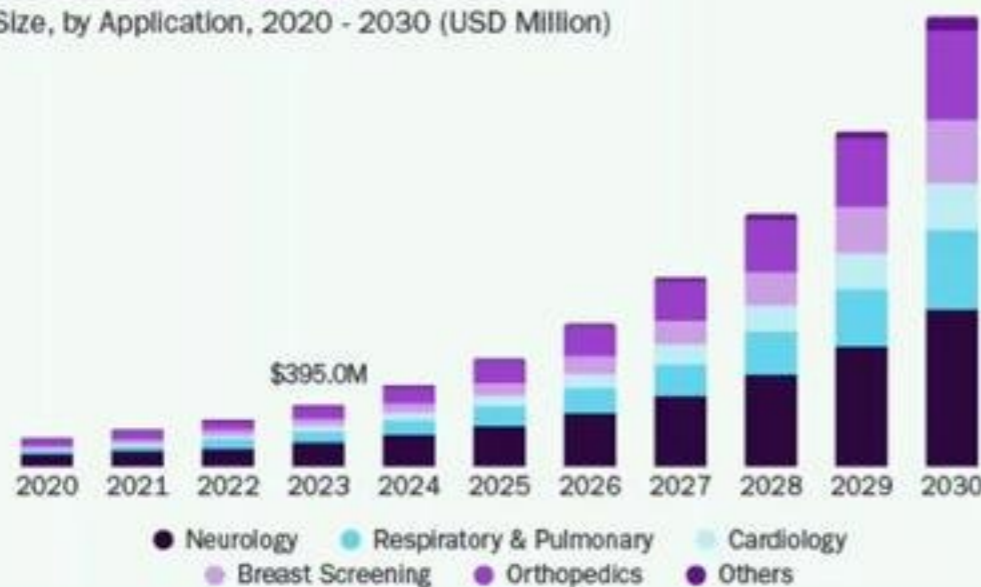
Exponential Rise

# Lane 1: AI for Clinical Decision Support

## Exponential Rise

### U.S. Artificial Intelligence (AI) in Medical Imaging Market

Size, by Application, 2020 - 2030 (USD Million)



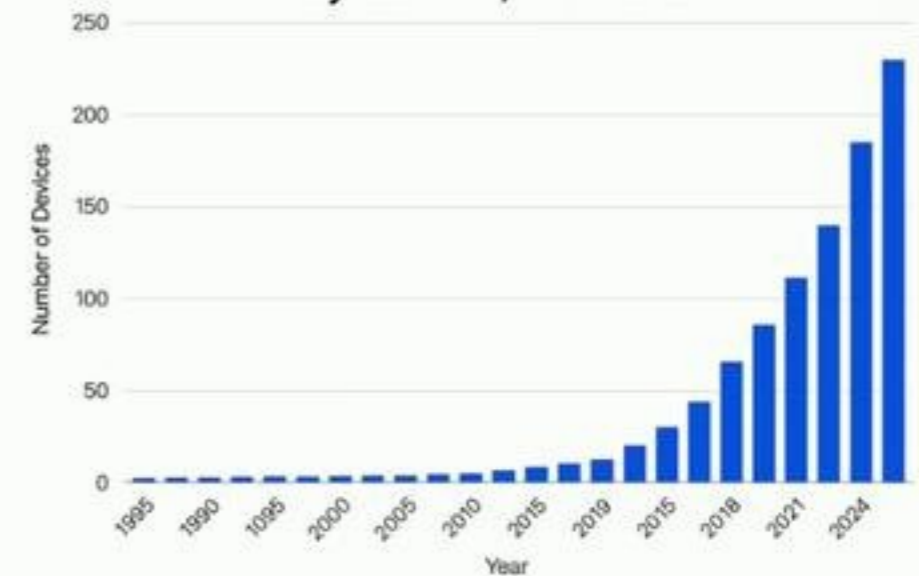
GRAND VIEW RESEARCH

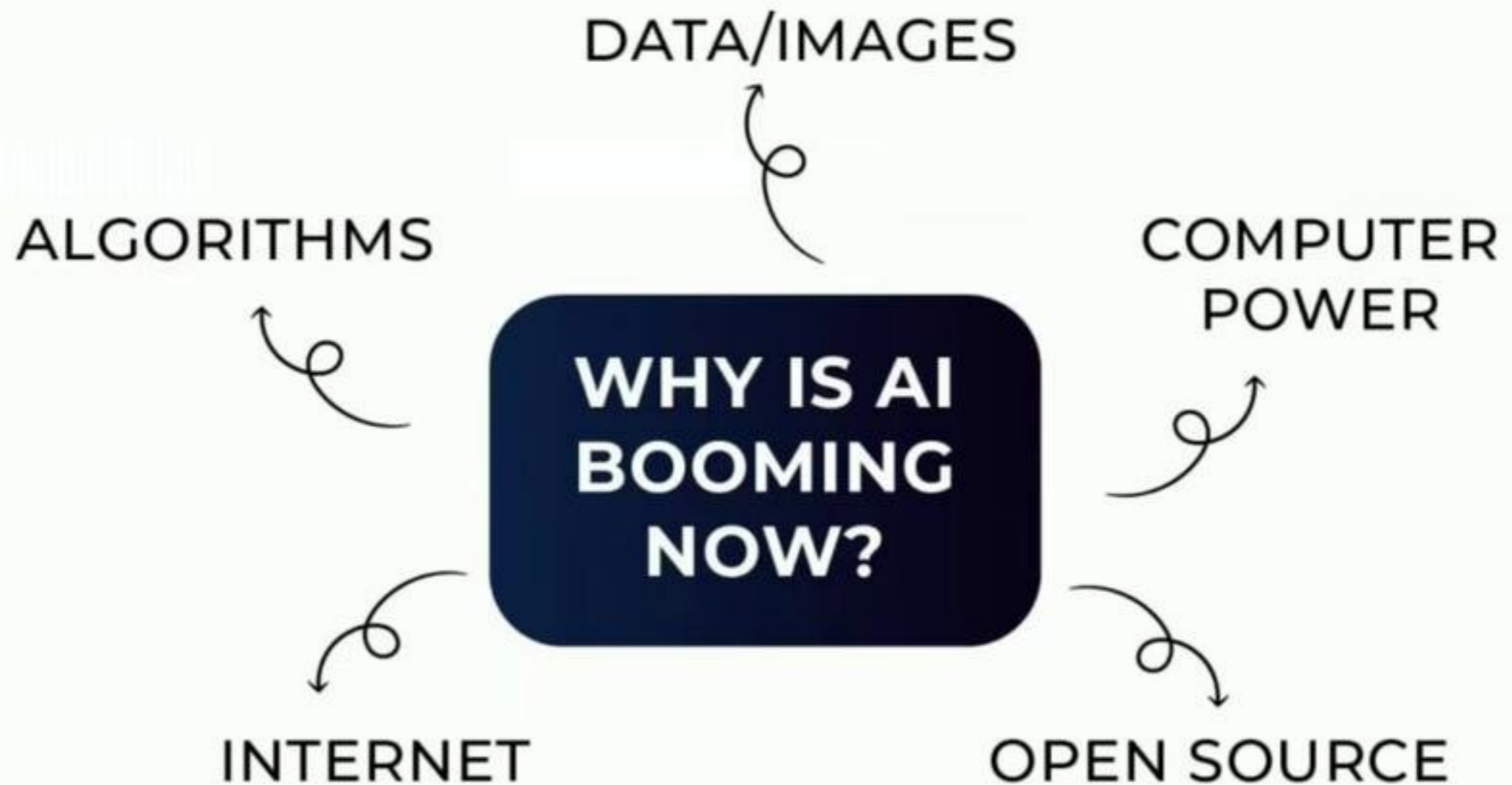
**33.2%**

U.S. Market CAGR,  
2024 - 2030

Source:  
www.grandviewresearch.com

### Growth of AI Medical Devices Authorized by the FDA, 1995-2024



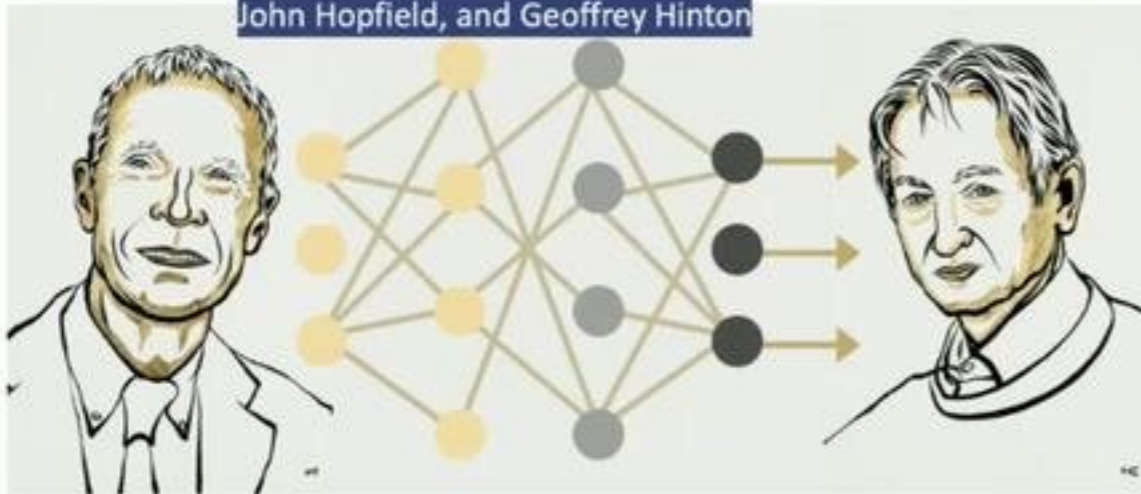


# How AI Works

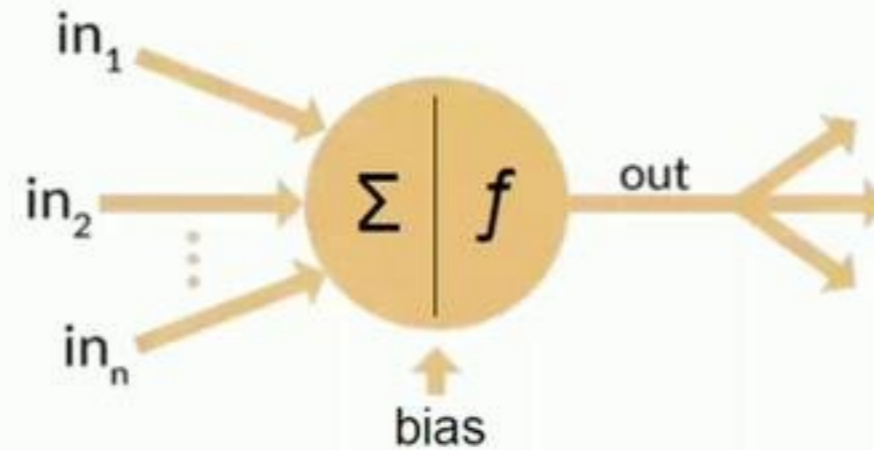
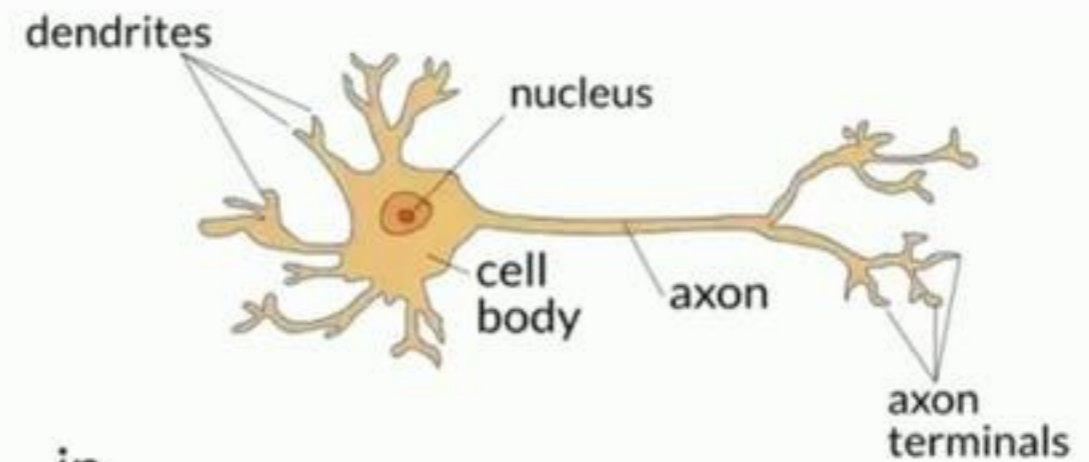
John Hopfield, and Geoffrey Hinton

# How AI Works

John Hopfield, and Geoffrey Hinton



## Neural Network Architecture



# Deep Learning

## ARTIFICIAL INTELLIGENCE

Any technique that enables computers to mimic human behavior



## MACHINE LEARNING

Ability to learn without explicitly being programmed

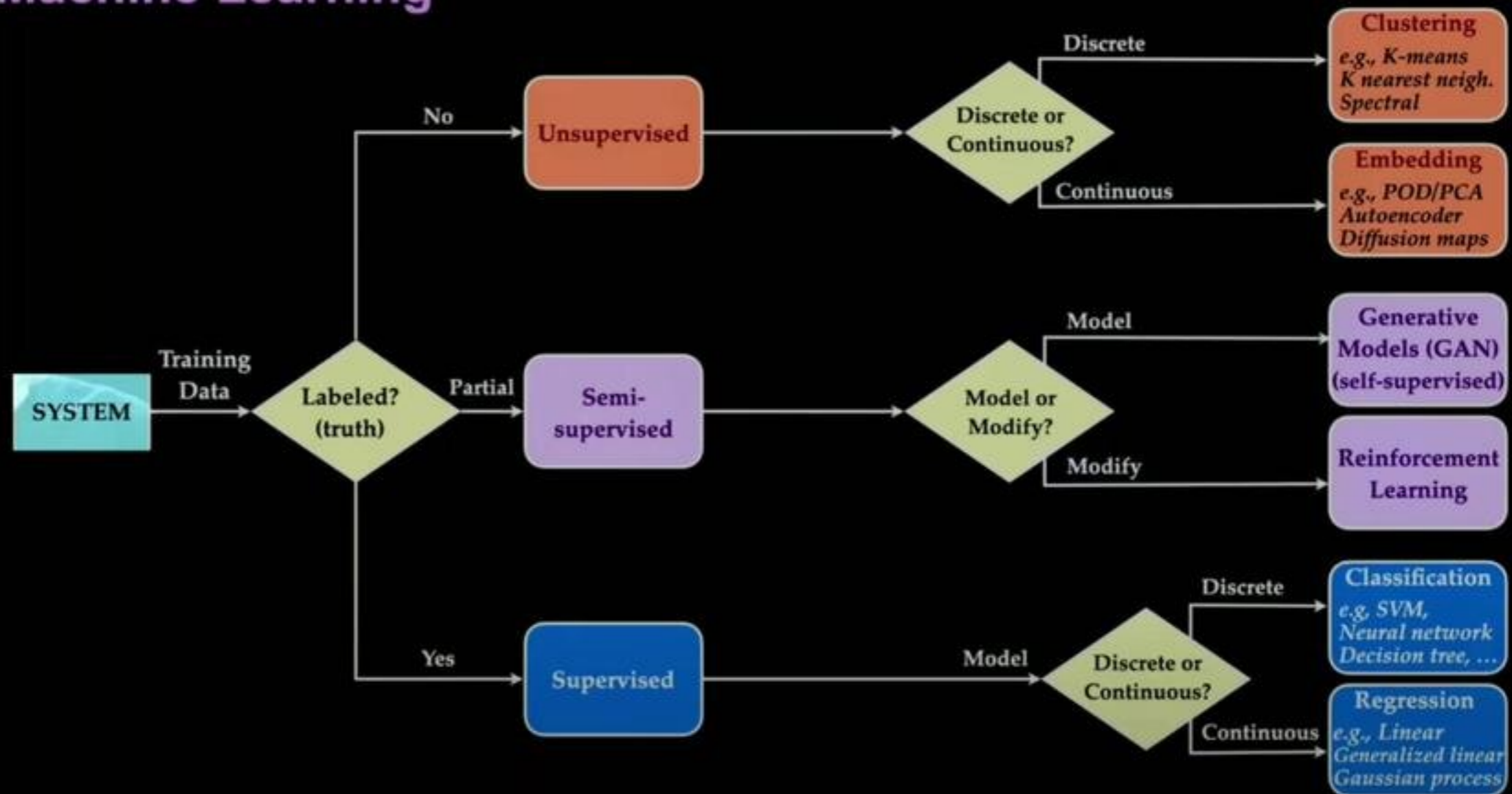


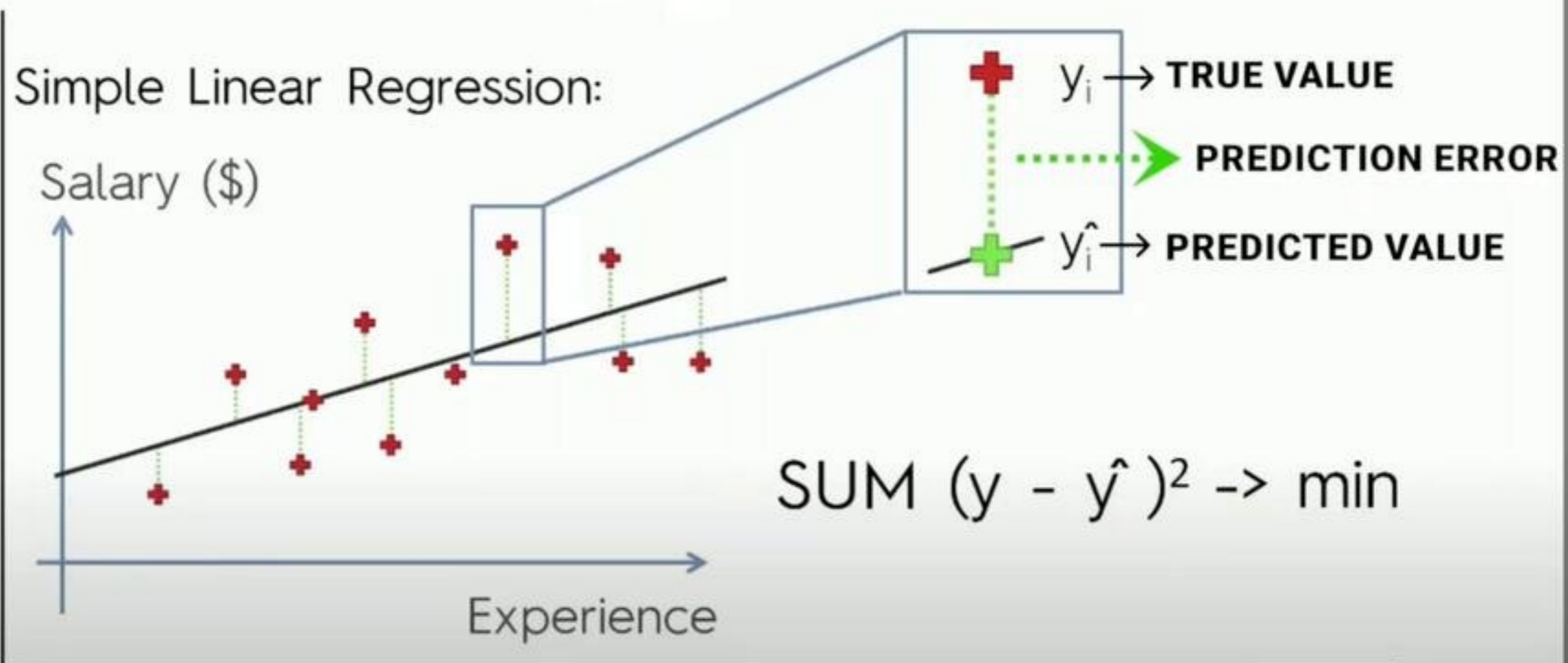
## DEEP LEARNING

Extract patterns from data using neural networks

3 1 3 4 7 2  
1 7 4 2 3 5

# Types of Machine Learning





## Simple Linear Regression:

Salary (\$)

Dependent Variable  
(Response Variable)

Independent Variables  
(Predictors)

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \varepsilon$$

Y intercept

Slope  
Coefficient

Error Term



$y_i$

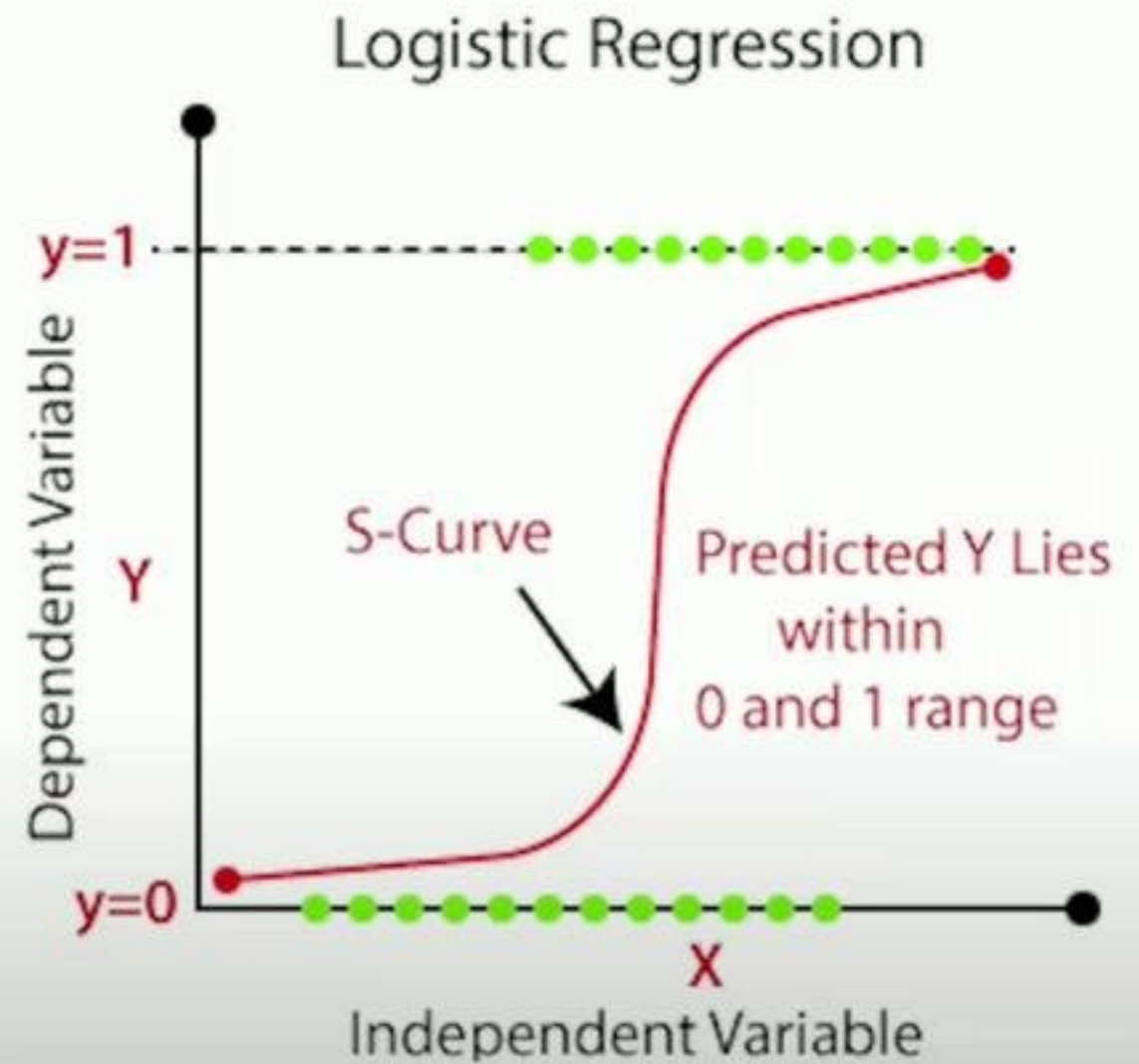
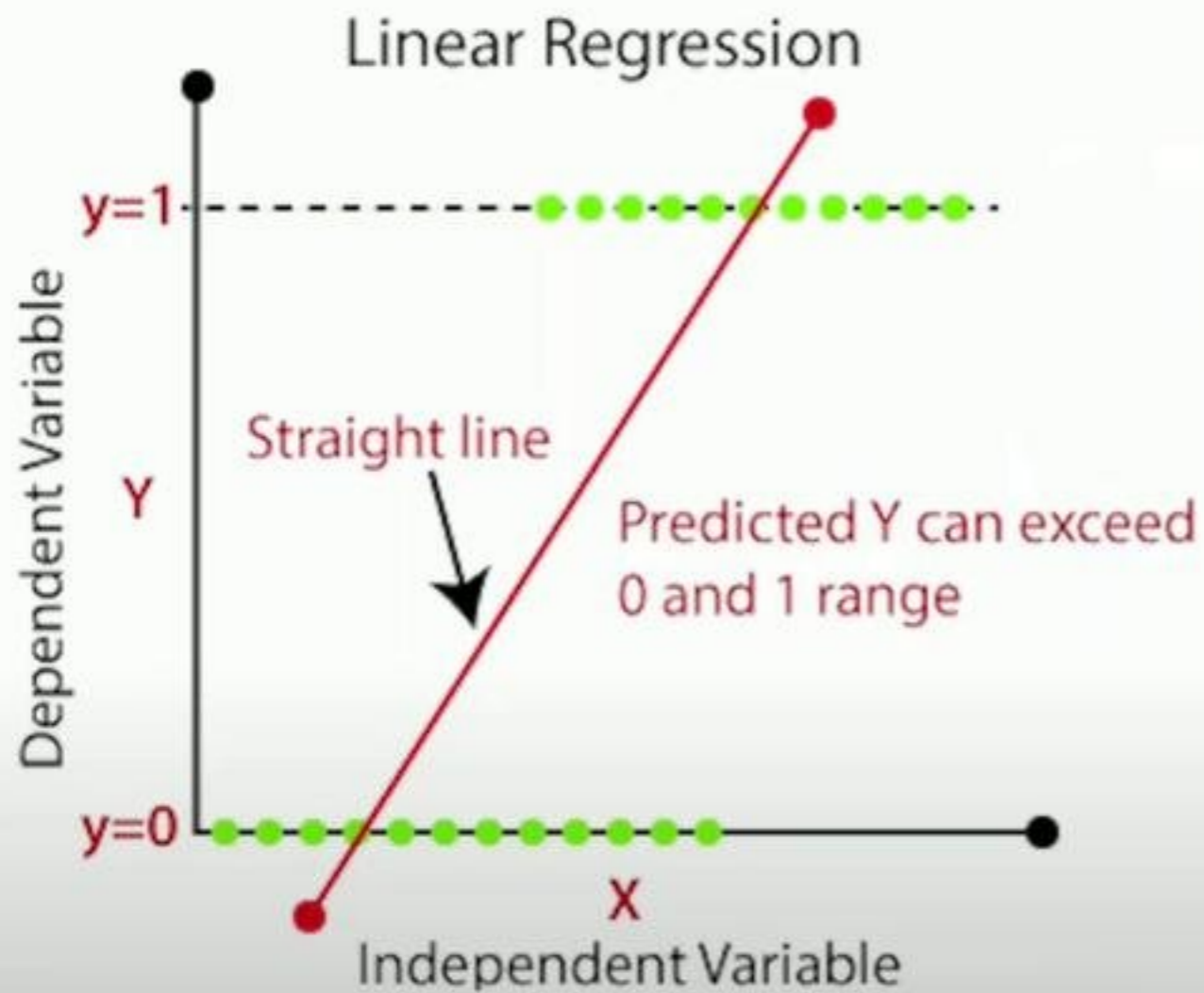
TRUE VALUE



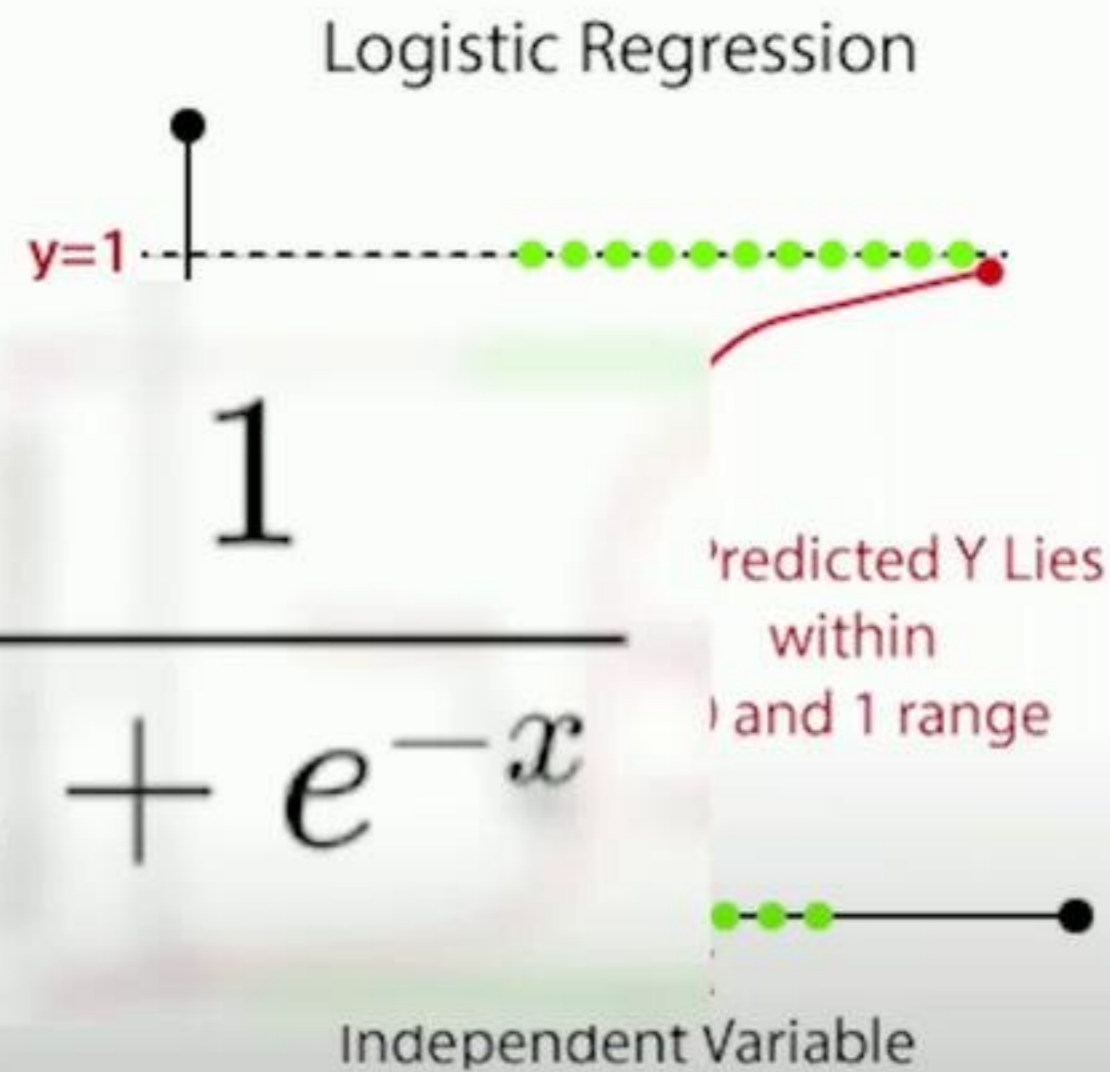
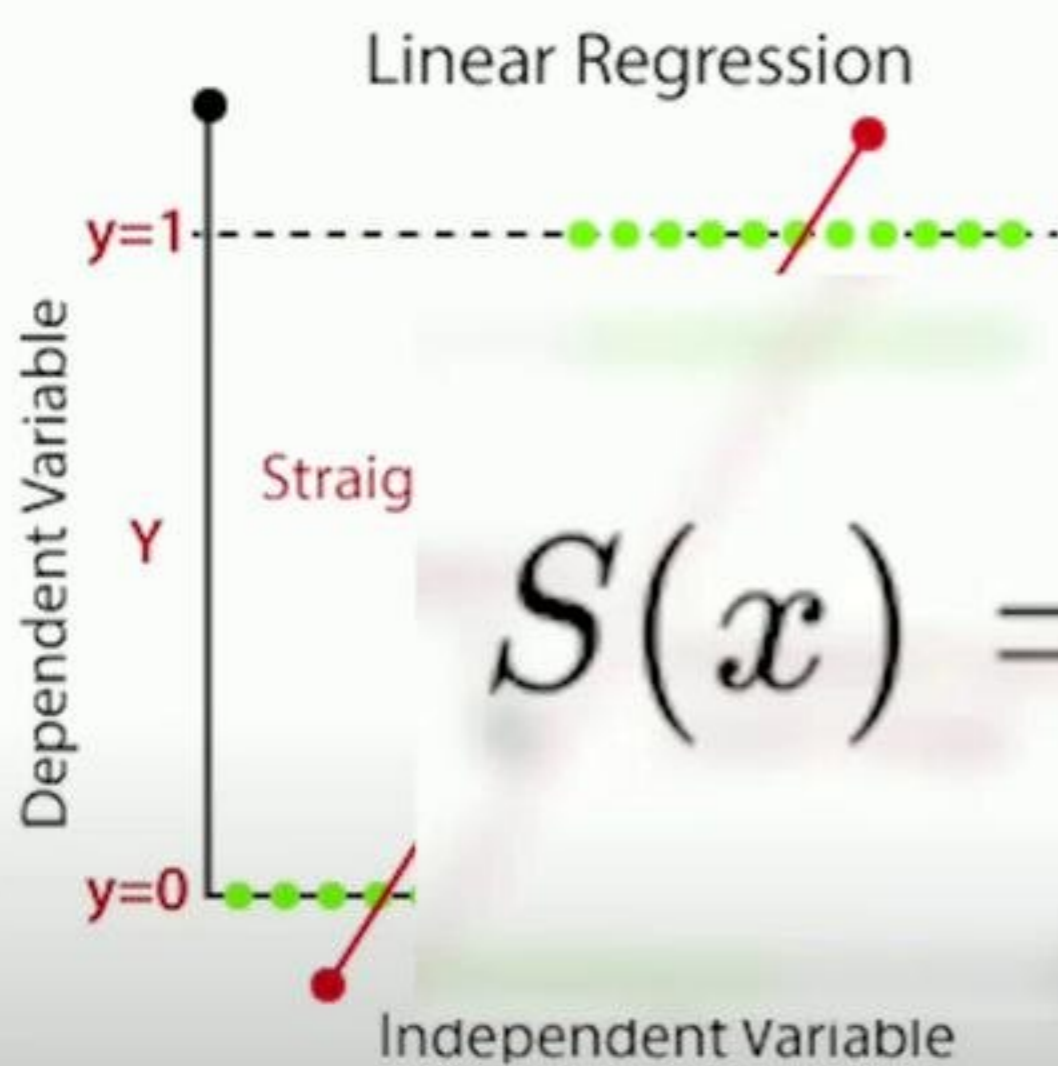
PREDICTION ERROR

PREDICTED VALUE

min



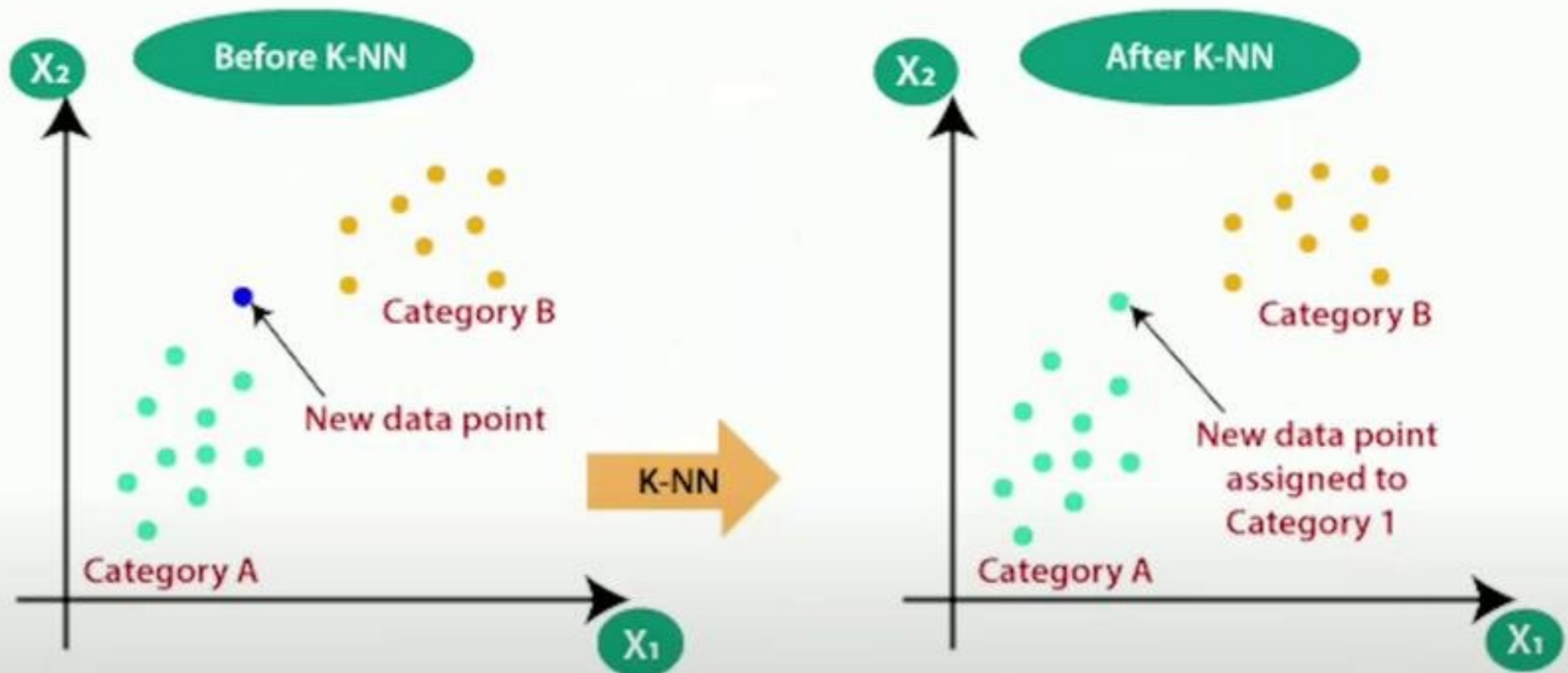
**SIGMOID FUNCTION**



$$S(x) = \frac{1}{1 + e^{-x}}$$

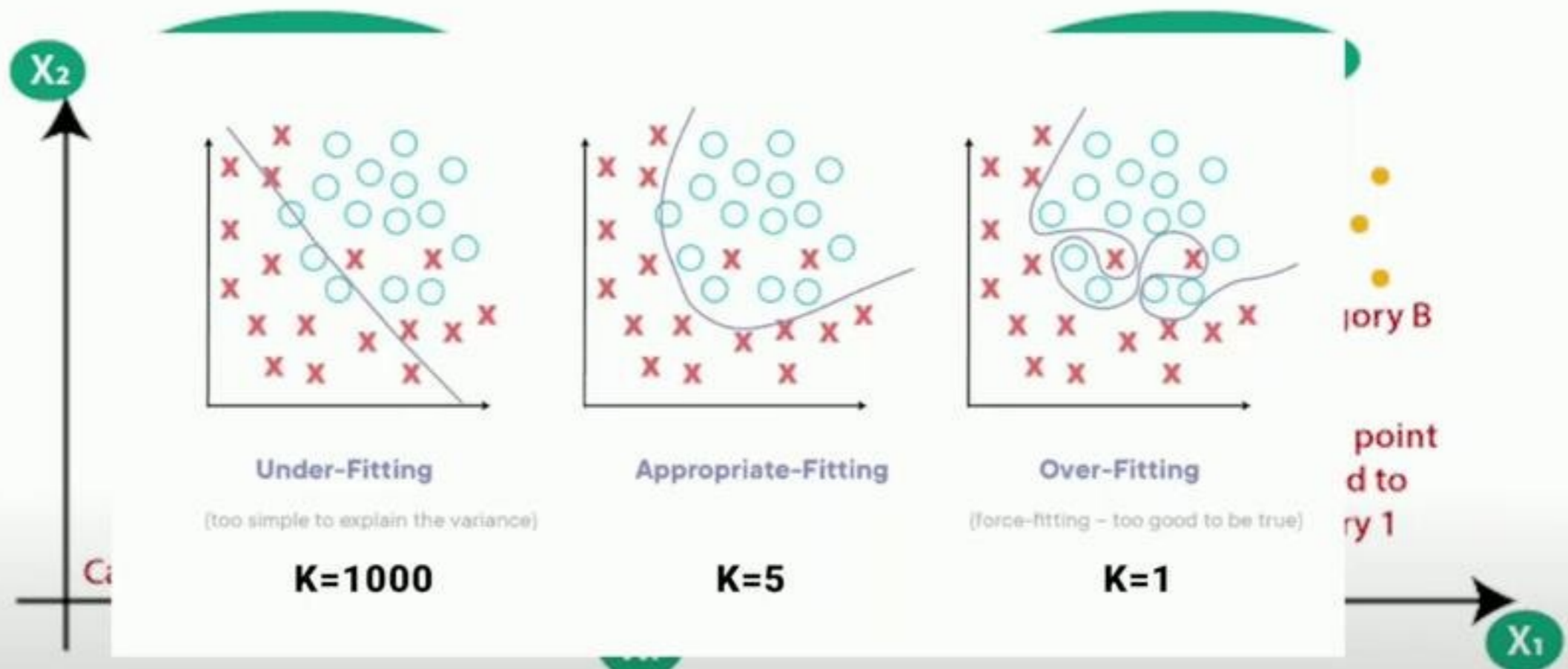
**SIGMOID FUNCTION**

## K NEAREST NEIGHBORS (KNN)



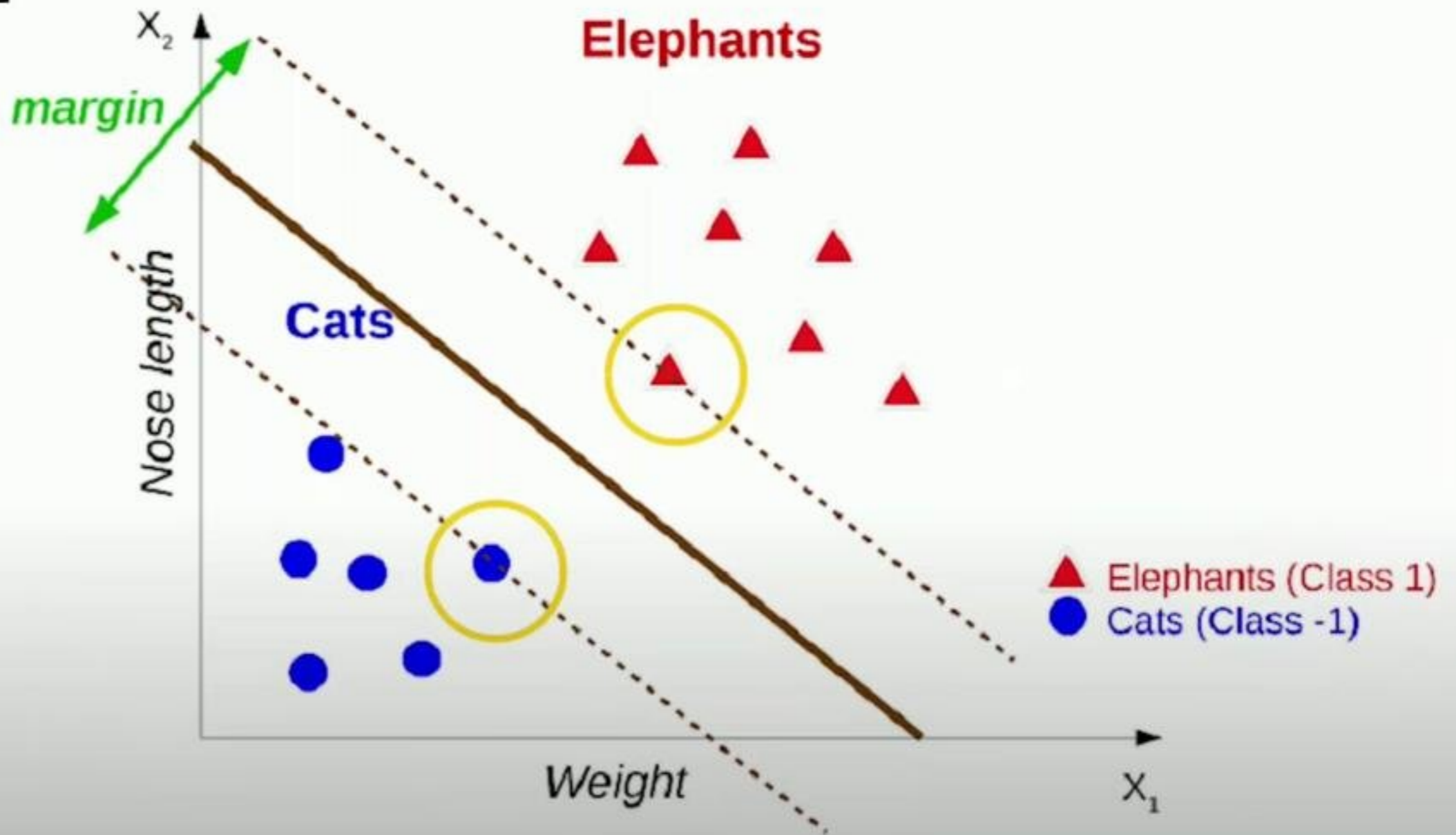
## NON-PARAMETRIC ALGORITHM

## K NEAREST NEIGHBORS (KNN)

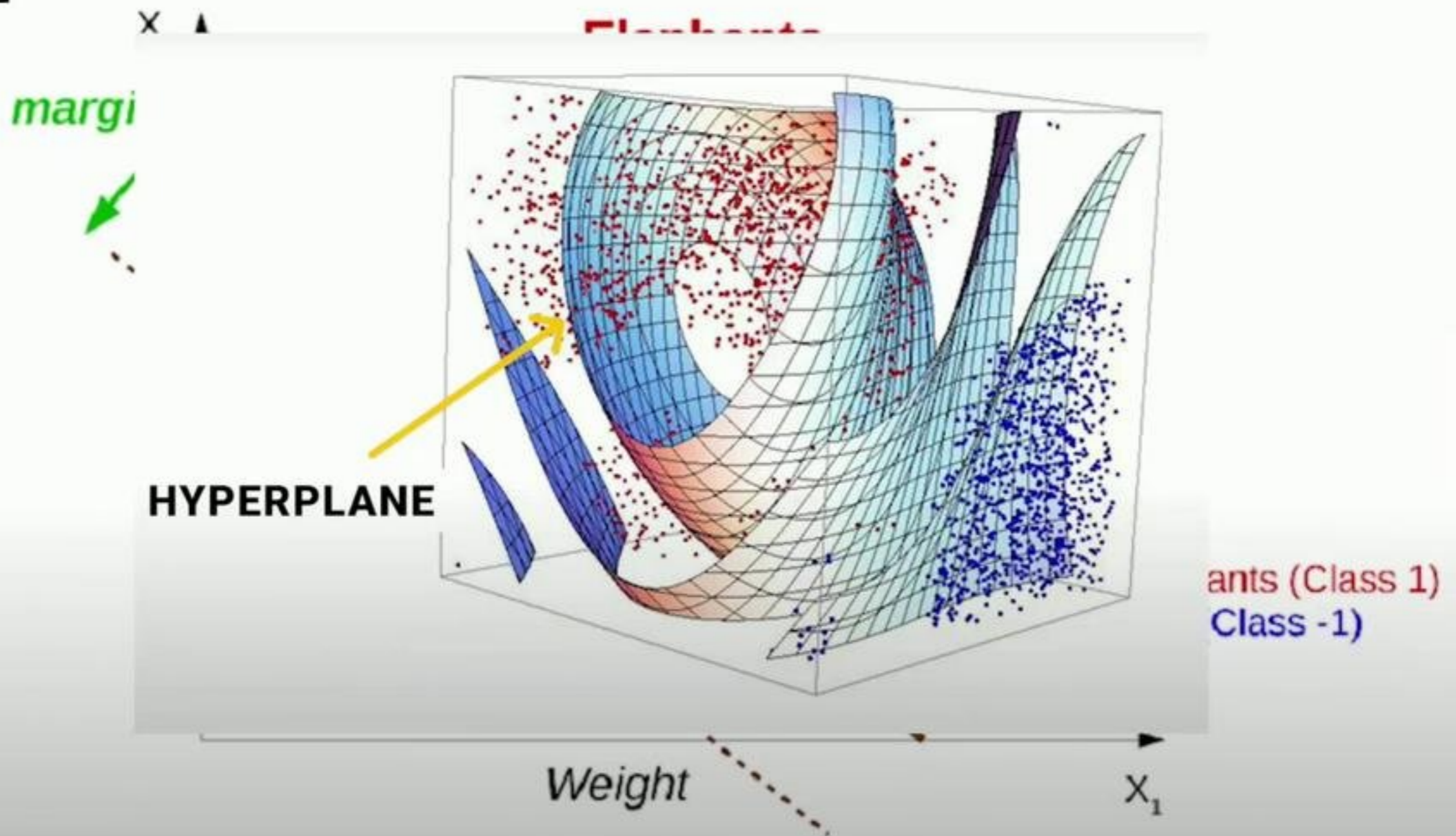


## NON-PARAMETRIC ALGORITHM

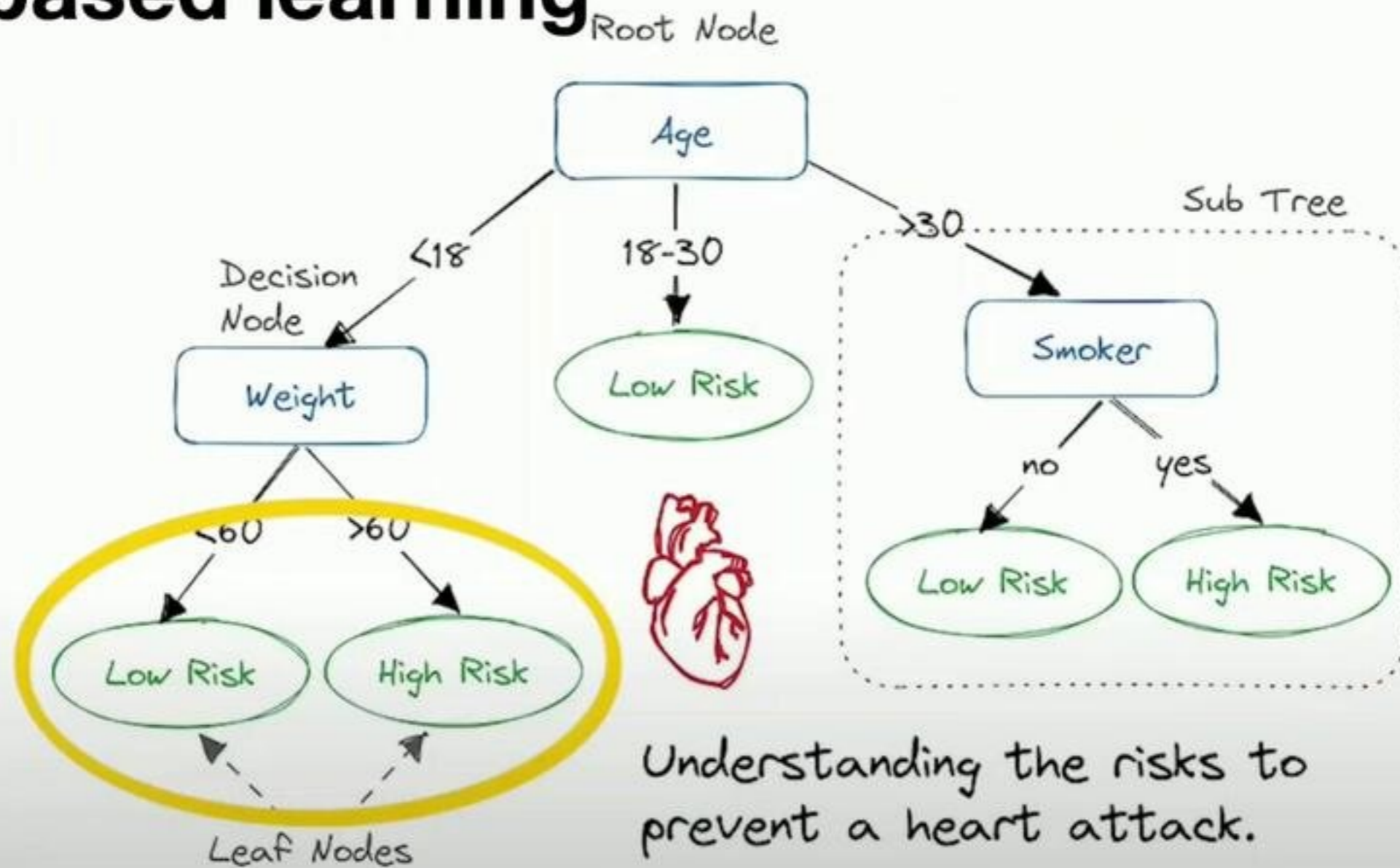
# SVM



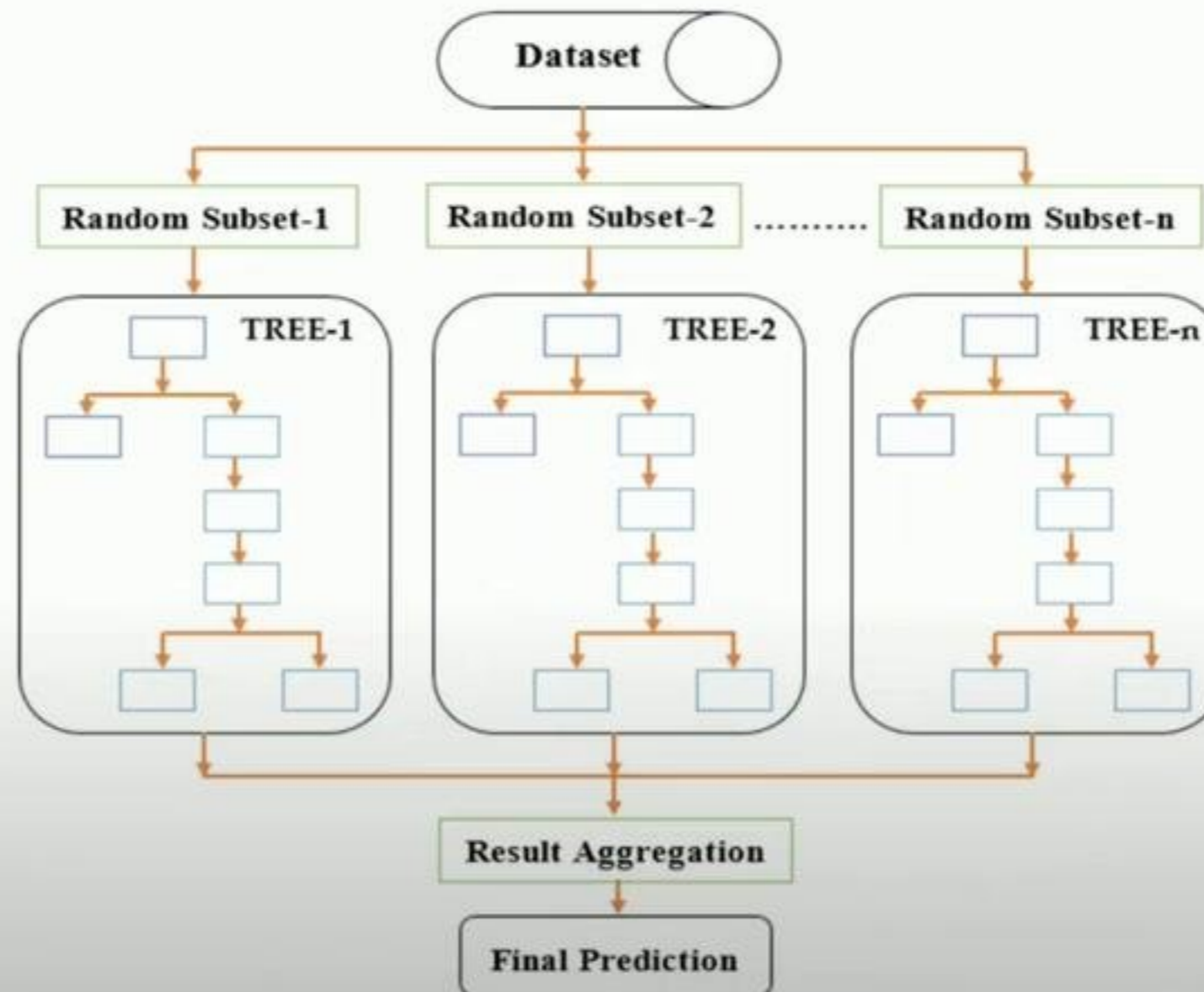
# SVM



# Trees based learning



# BAGGING



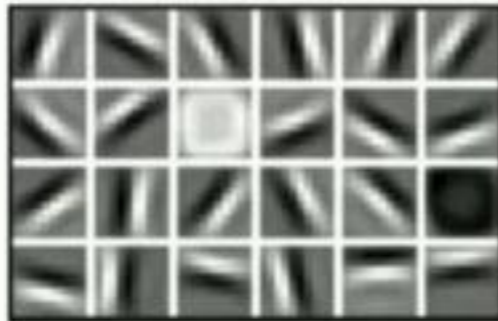
# Deep Learning

Progressive extraction of features by a Neural Network

Input image



Low Level Features



Lines & Edges

Mid Level Features

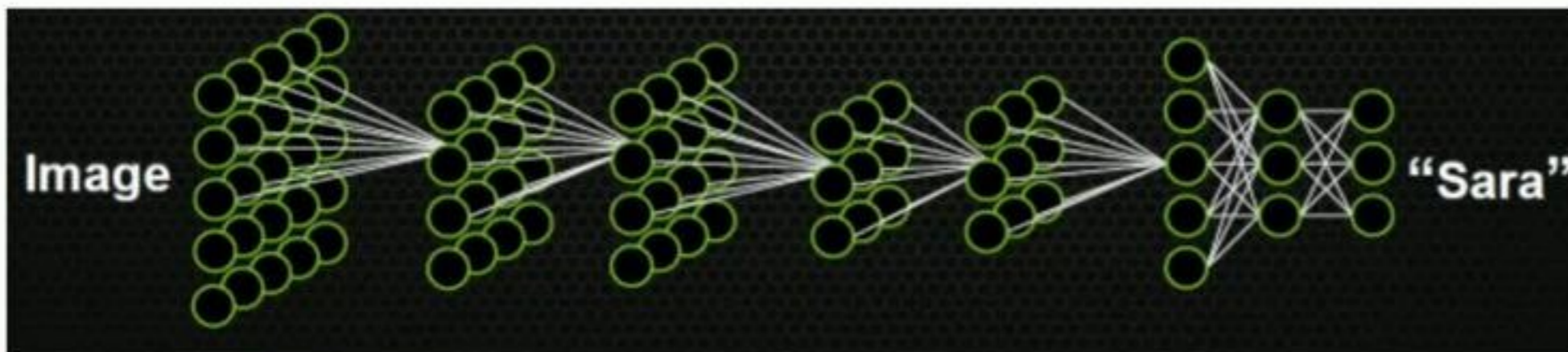


Eyes & Nose & Ears

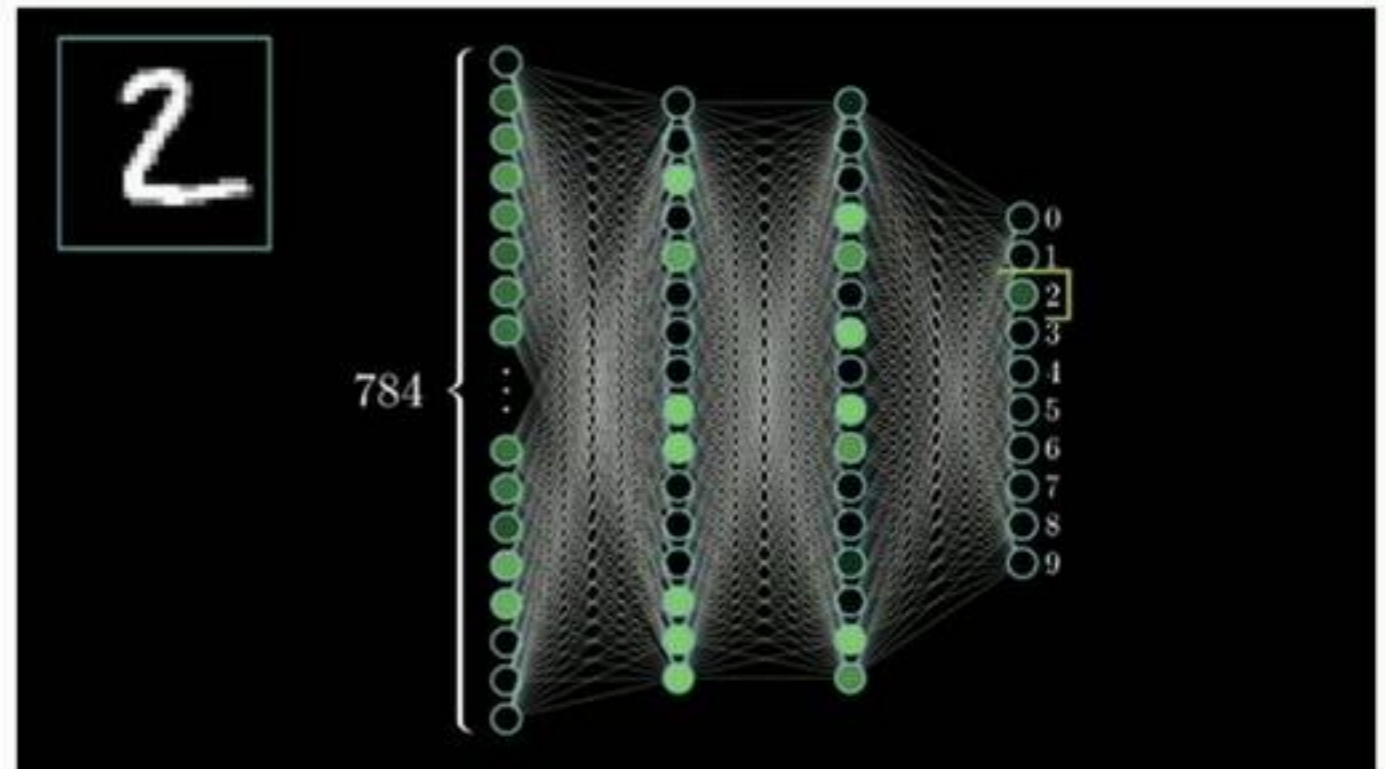
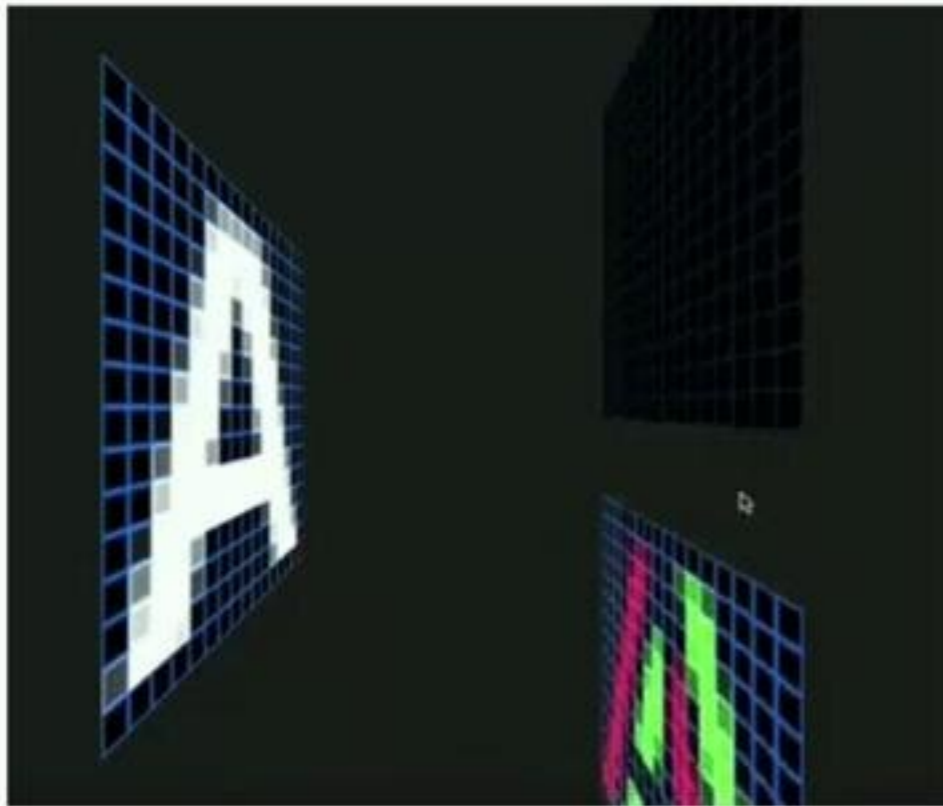
High Level Features



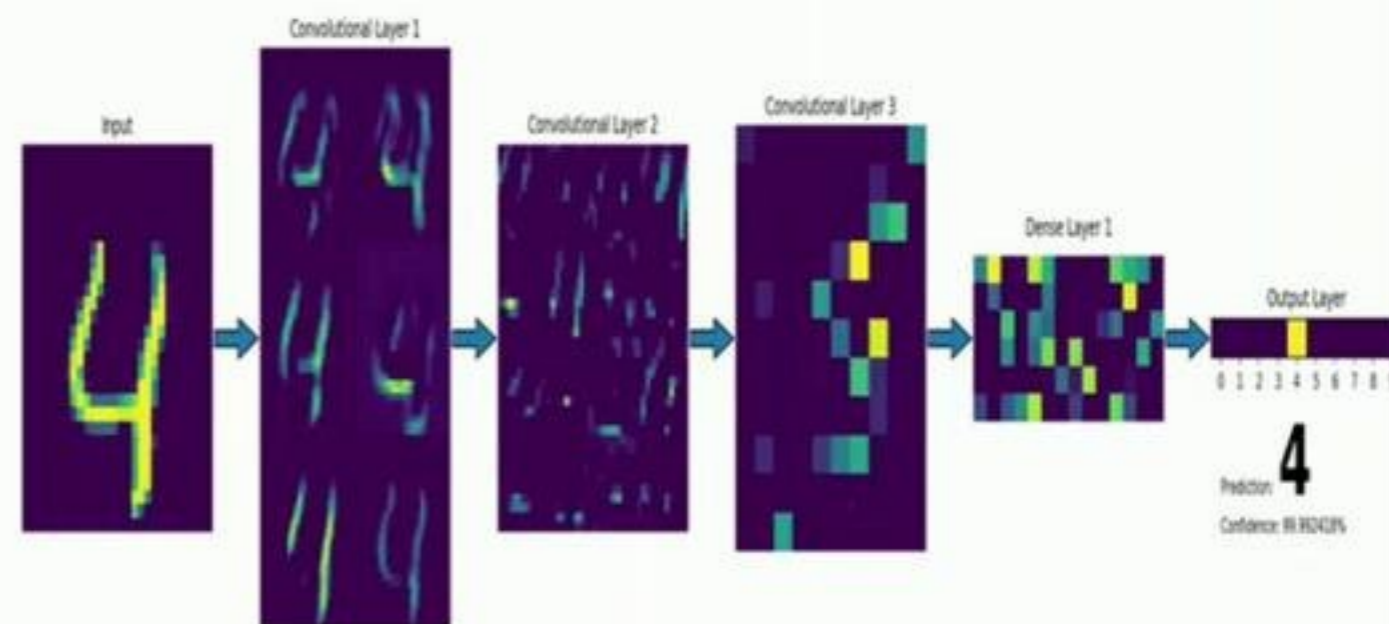
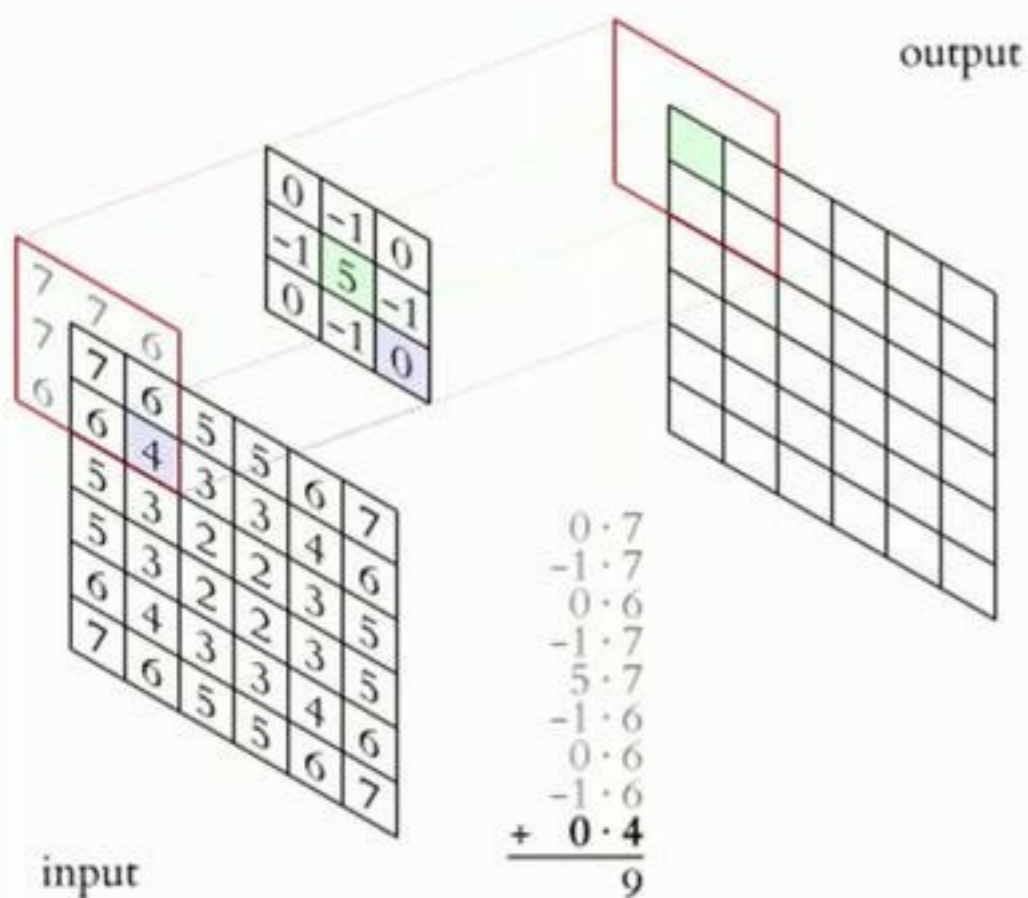
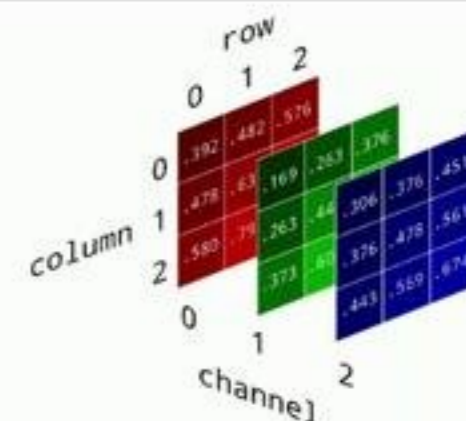
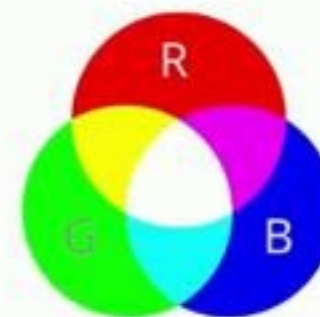
Facial Structure

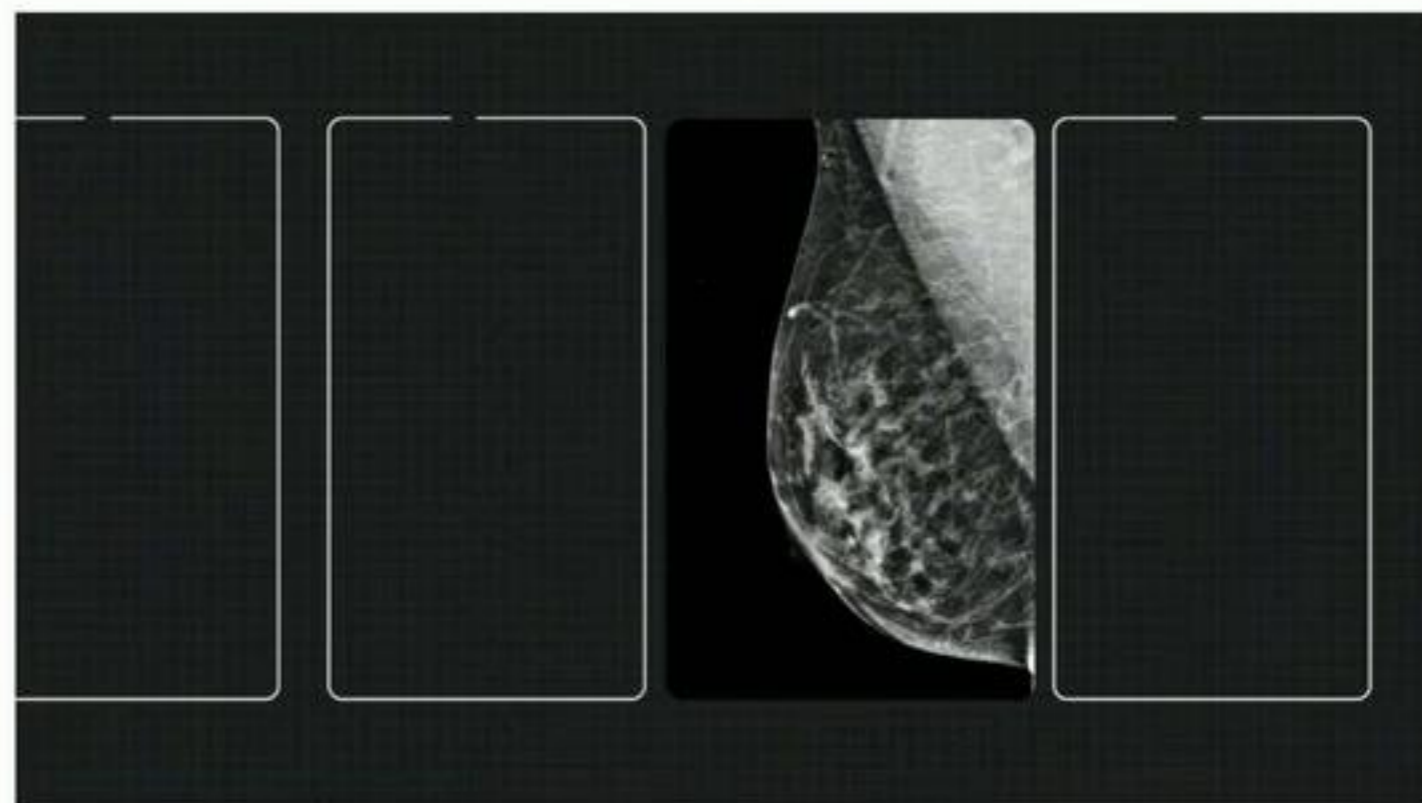


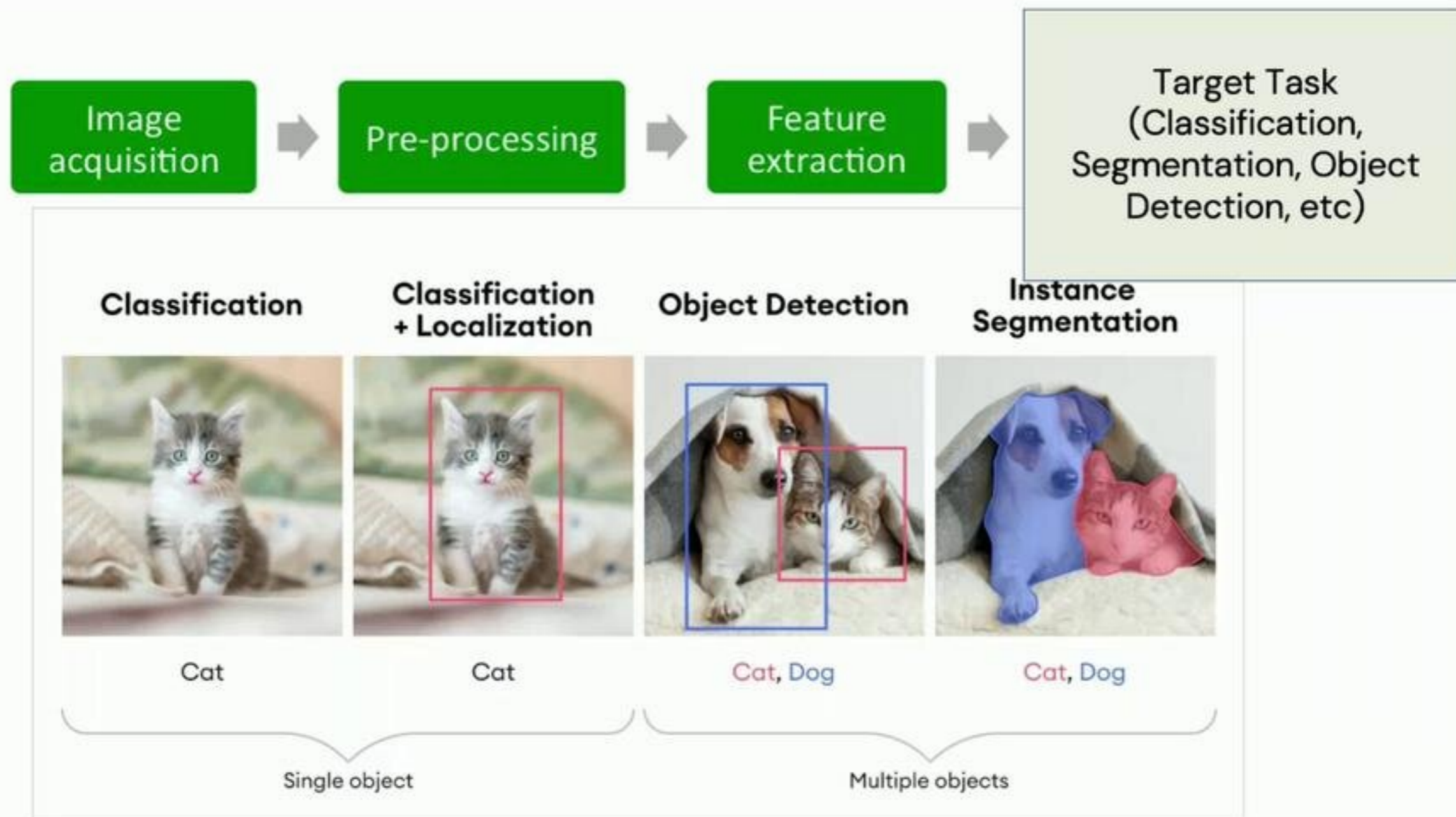
# How Deep Learning Works



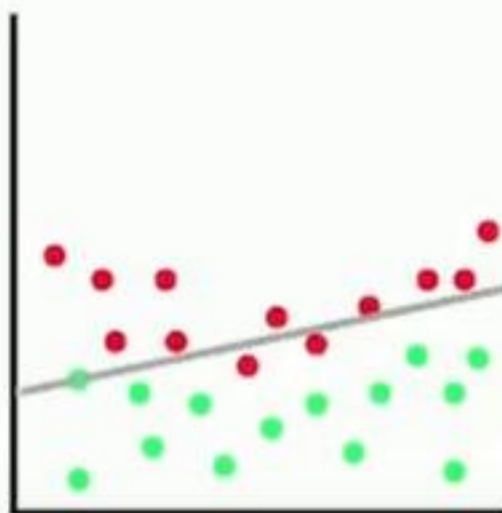
# How Deep Learning Works



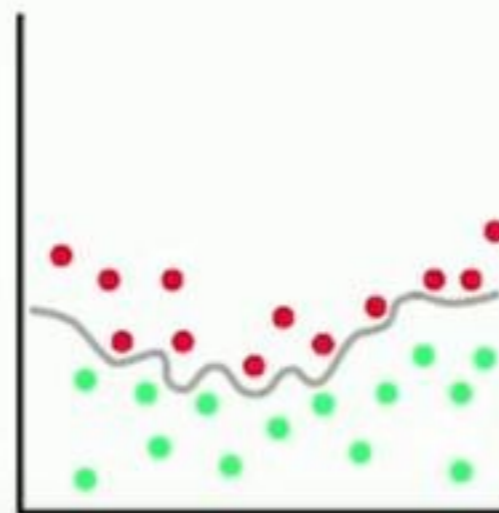




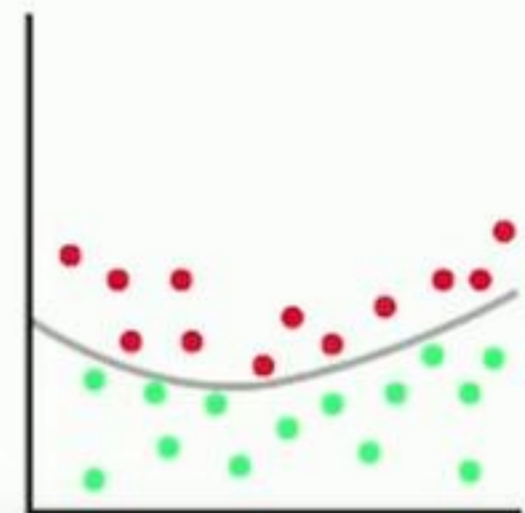
learning & regularization



Underfitting

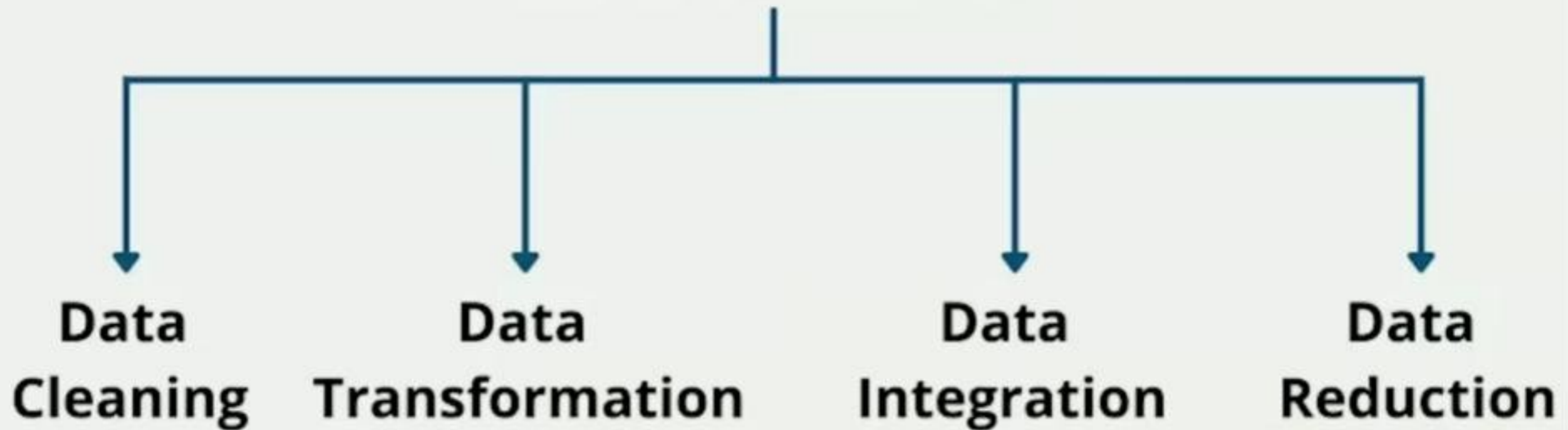


Overfitting



Balanced

# Data Preprocessing



- Removing Duplicates
- Handling missing values

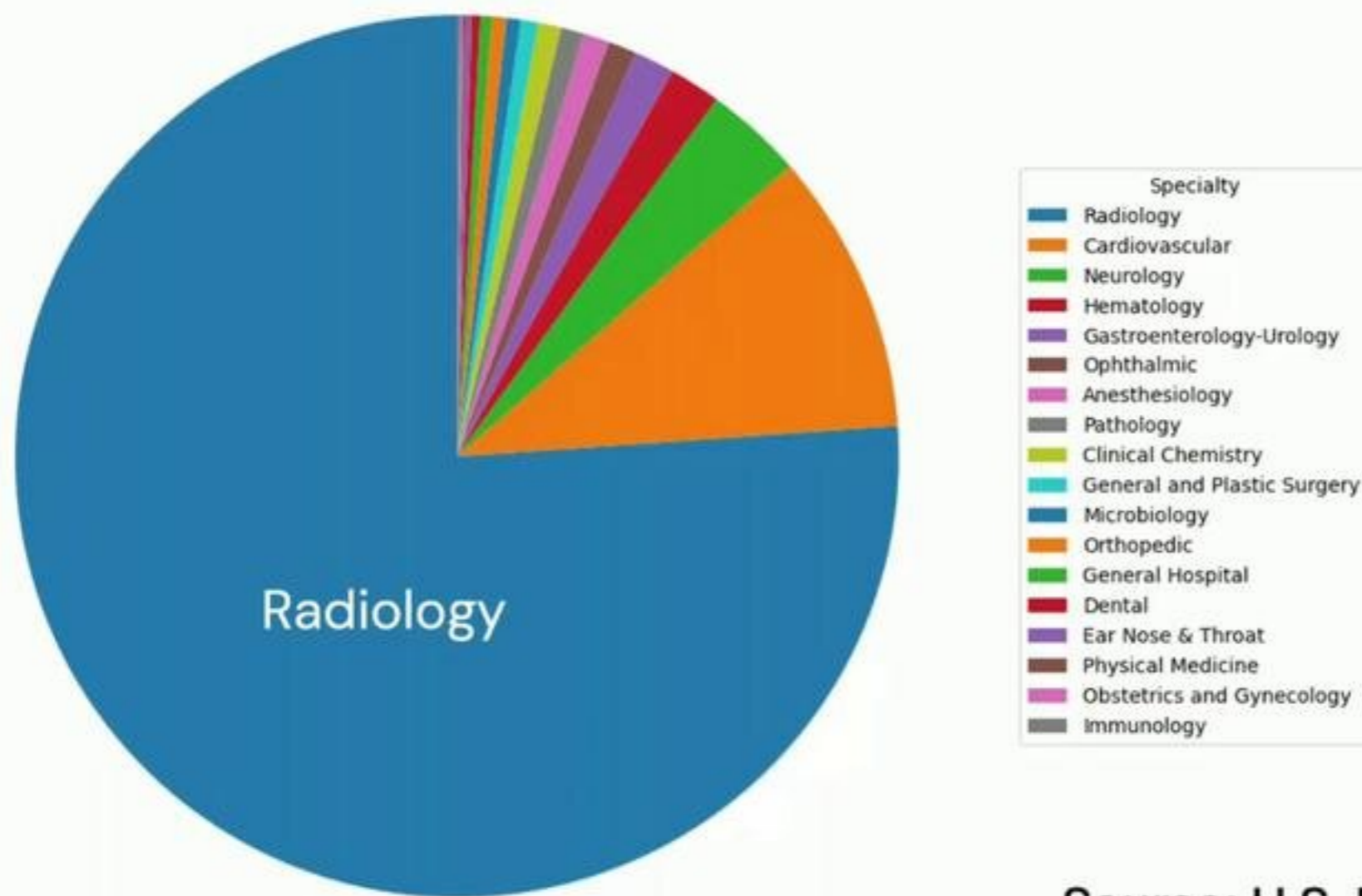
- Scaling
- Encoding

- Joining
- Merging

- Sampling
- Dimensionality Reduction

<https://www.youtube.com/watch?v=E0Hmnixke2g>

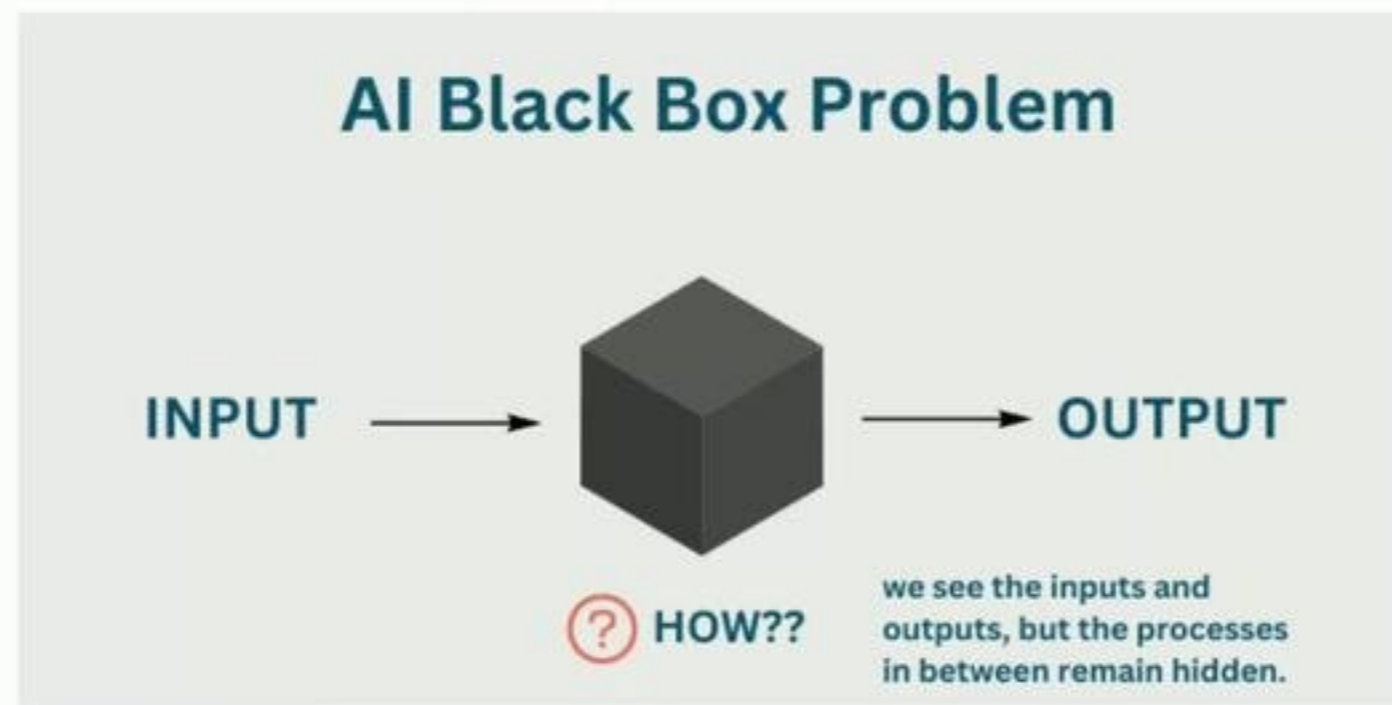
# Landscape – FDA-approved AI cases by specialty



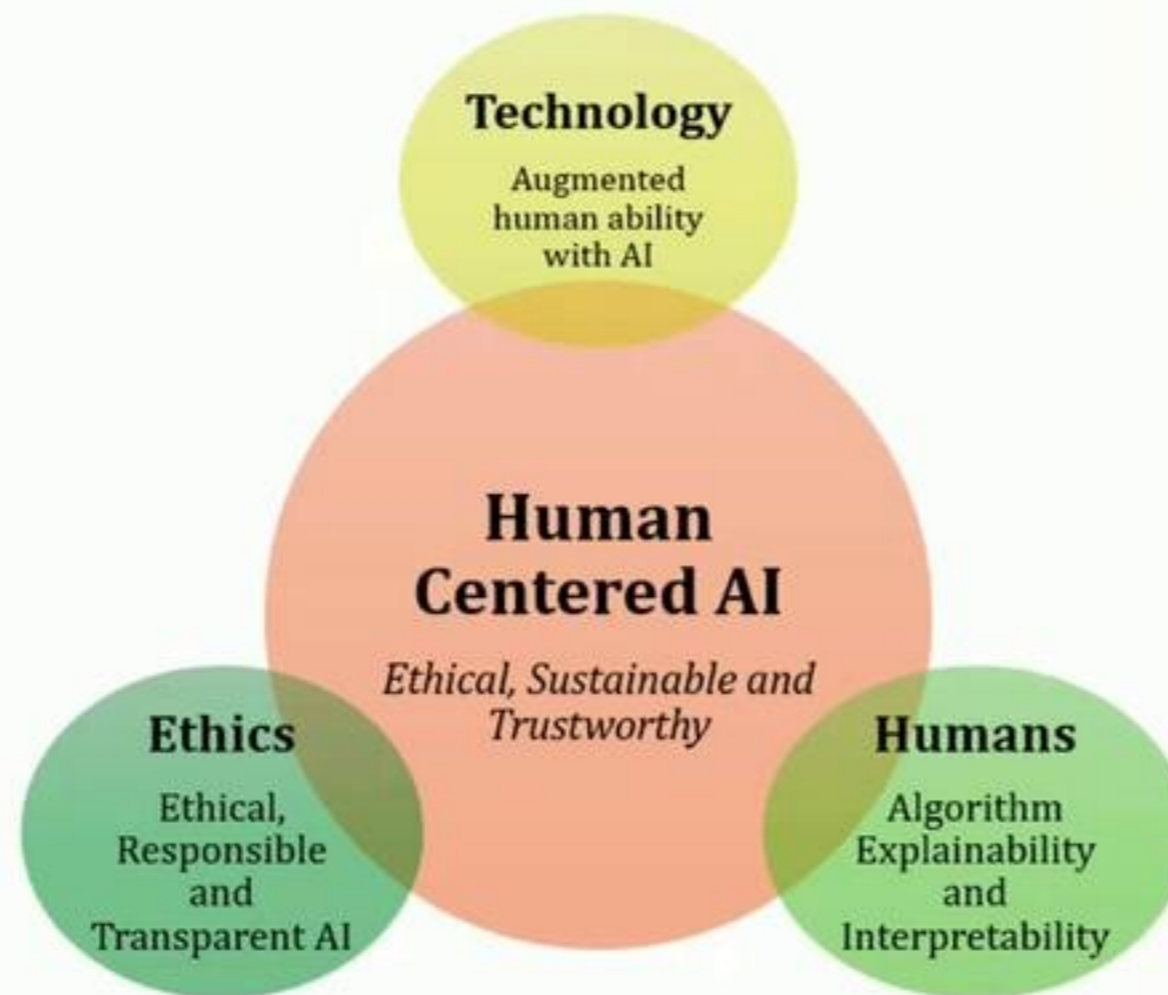
Source: U.S. Food & Drug Administration

# Rise of Black-Box Clinical AI

- Deep CNNs for imaging, transformers for EHR – high accuracy.
- Opaque internals raise concerns over failure modes & bias.
- Human-Centered and Explainable AI (XAI) bridges gap between performance and accountability.

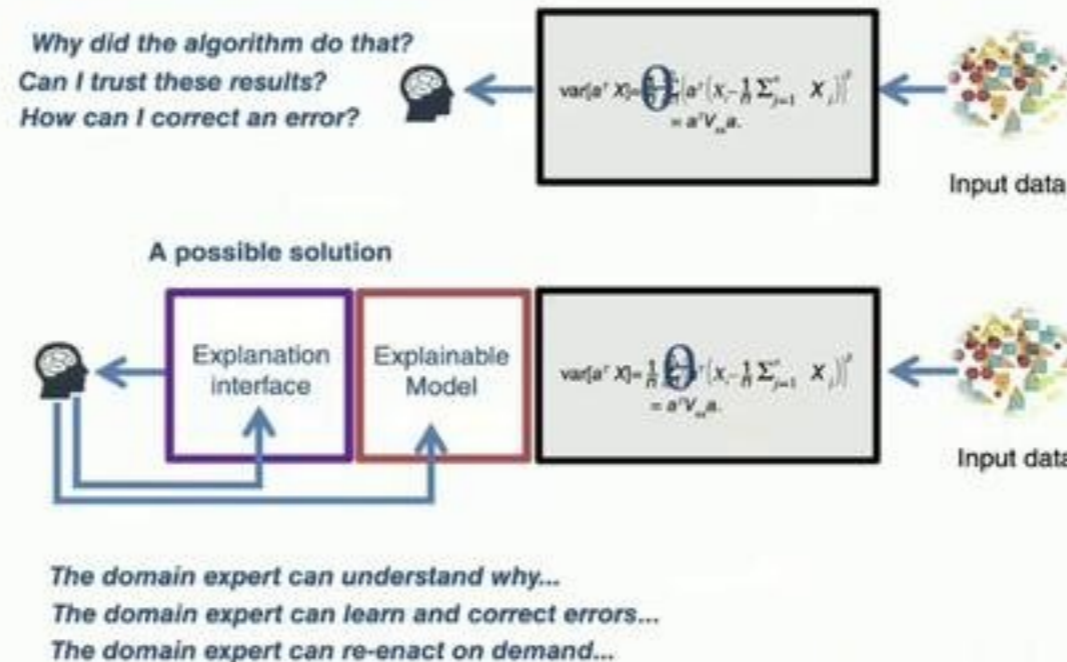


# Human Centered AI (HCAI)



# What is Explainable AI (XAI)?

- Narrow: methods that make model decisions understandable.
- Broad: anything making AI systems transparent (data, performance).

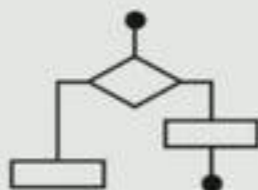


# xAI and Human-Centered AI is more than Models

## AI Solution



Features



Algorithm



Model Parameters



Model

Each element constituent of the solution process needs to be explainable for the solution to be truly explainable [Lipton 2016]

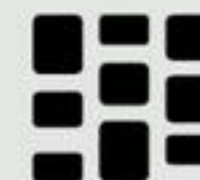
## User



Cognitive Capacity



Domain Knowledge



Explanation Granularity

# Why HCAI and XAI is Critical in Medicine

- High-stakes decisions: diagnosis, triage, treatment planning.
- Regulatory pressure: GDPR 'right to explanation', FDA SaMD draft.
- Clinician trust & patient safety depend on understandable models.

# Explain



Why did this model make a prediction?

# Interpret



How does this model work?

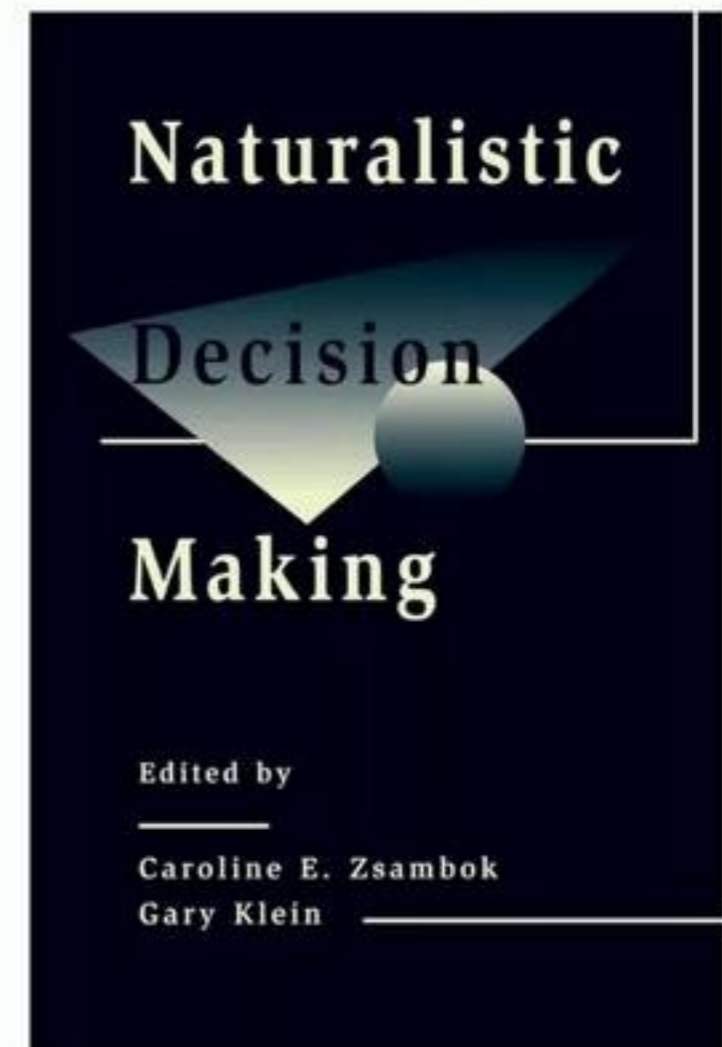
# Understand



What are the model's capabilities and limitations?

# Human Naturalistic Decision Making (NDM)

- NDM research to study how people make decisions in real-world settings.
- The NDM framework emphasizes the role of experience in enabling people to categorize situations to make effective decisions.



# How human make **decision**?

## System 1




Fast, intuitive and emotional

## System 2



Slow, conscious and effortful



THINKING,  
FAST AND SLOW  
  
DANIEL  
KAHNEMAN

## Recognition-Primed Decision Model (RPD)

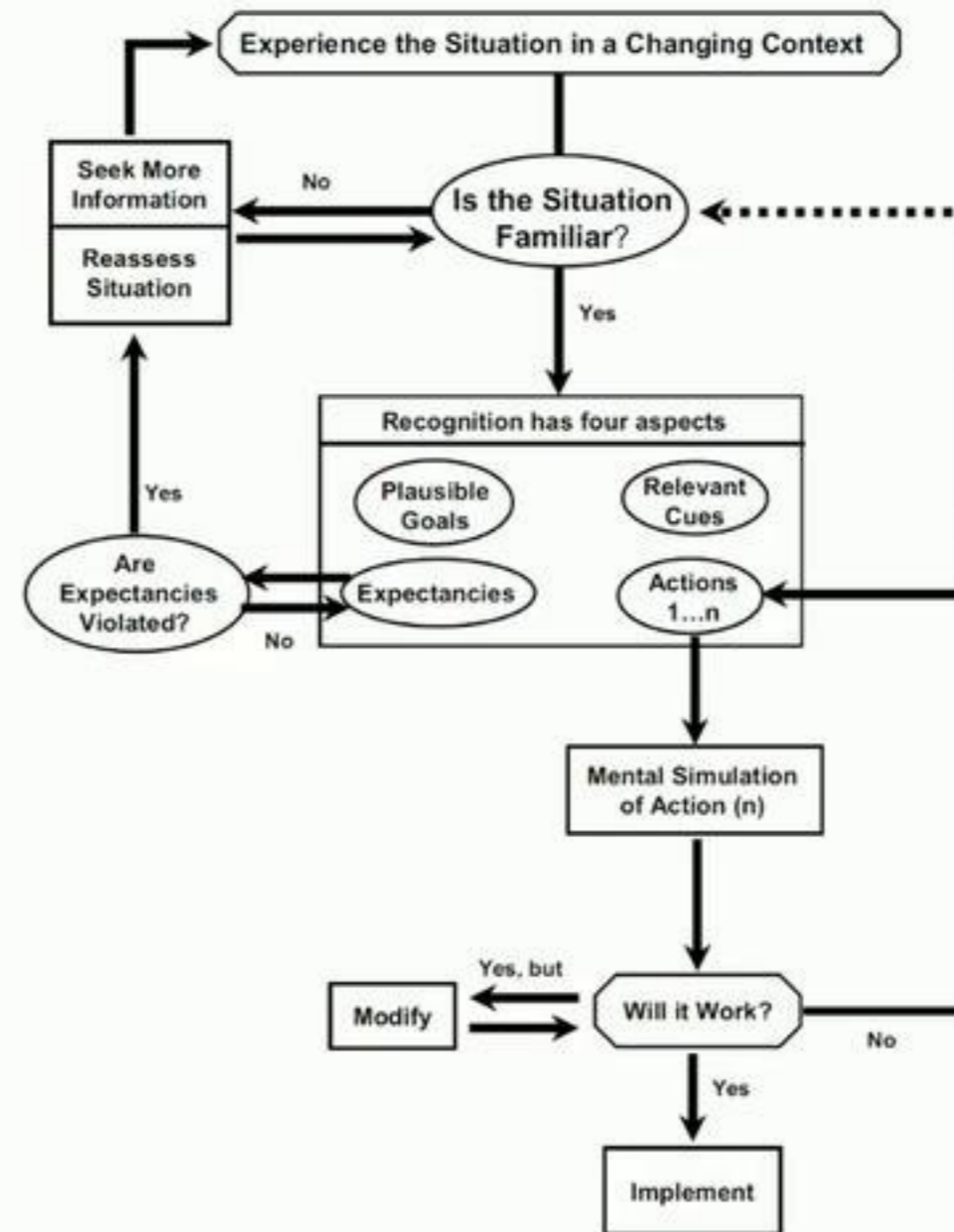
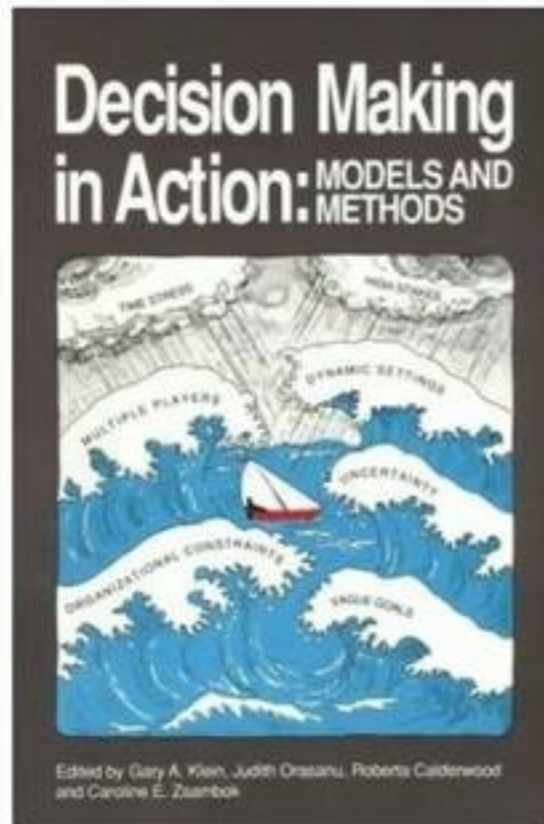
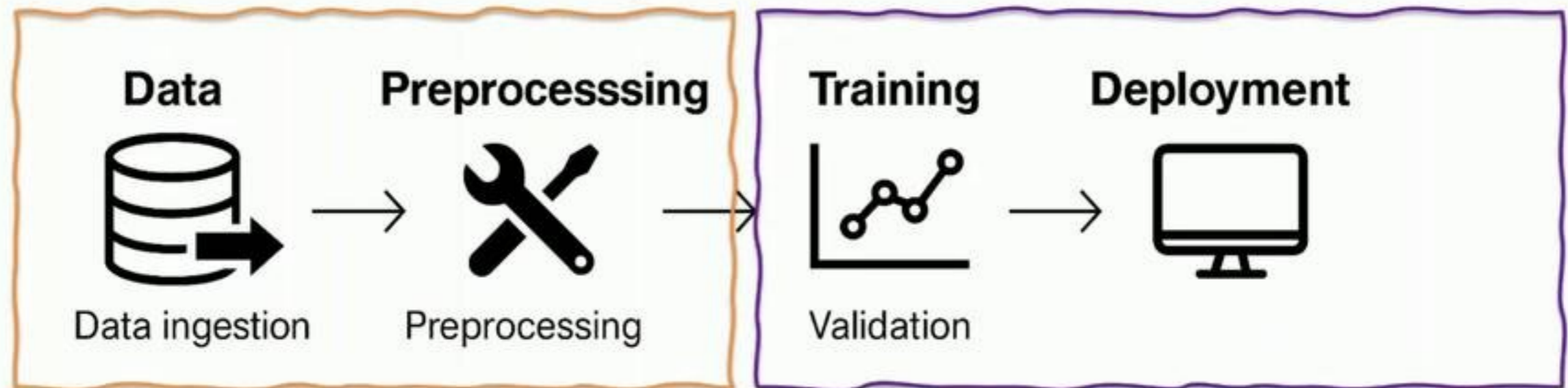


Figure 1. Model of recognition-primed decision making. (*Decision making in action: Models and methods*. G. A. Klein, J. Orasanu, R. Calderwood, C. E. Zsombok, Editors. Copyright © 1993 by Ablex Publishing Corporation, Norwood, NJ. Reproduced with permission of Greenwood Publishing Group, Inc., Westport, CT.)

# Where Explanations Fit in the Pipeline



- Pre-deployment: model debugging & bias checks

- At runtime: clinician-facing insights for each patient
- Post-deployment monitoring: drift & fairness audits

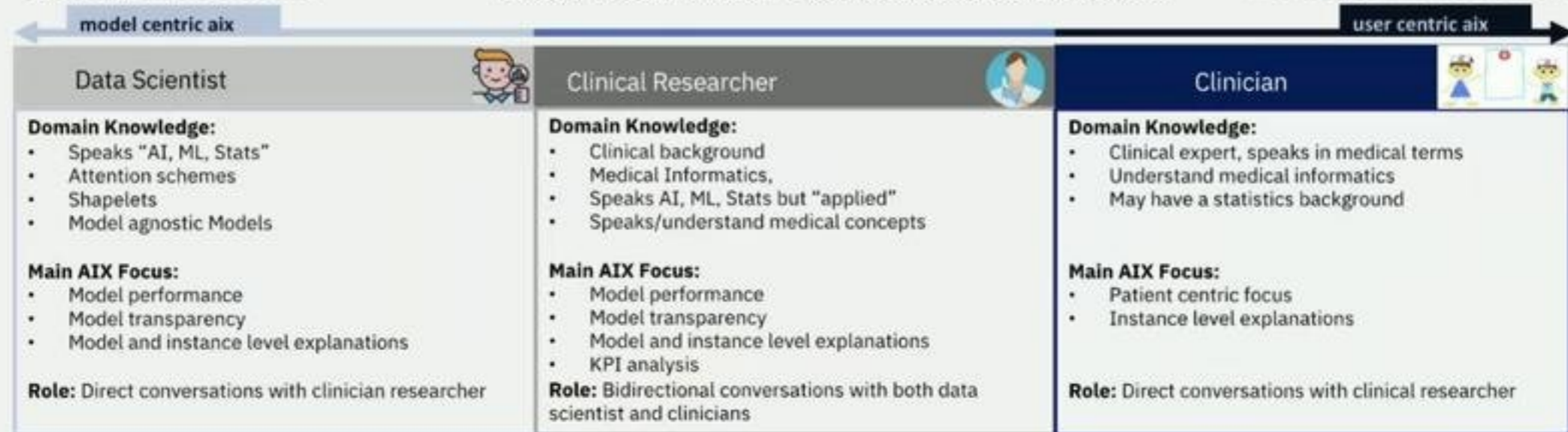
# Personas!



Must match the complexity & capability of the consumer

## The HealthXAI Persona Continuum

Must match the domain knowledge of the consumers



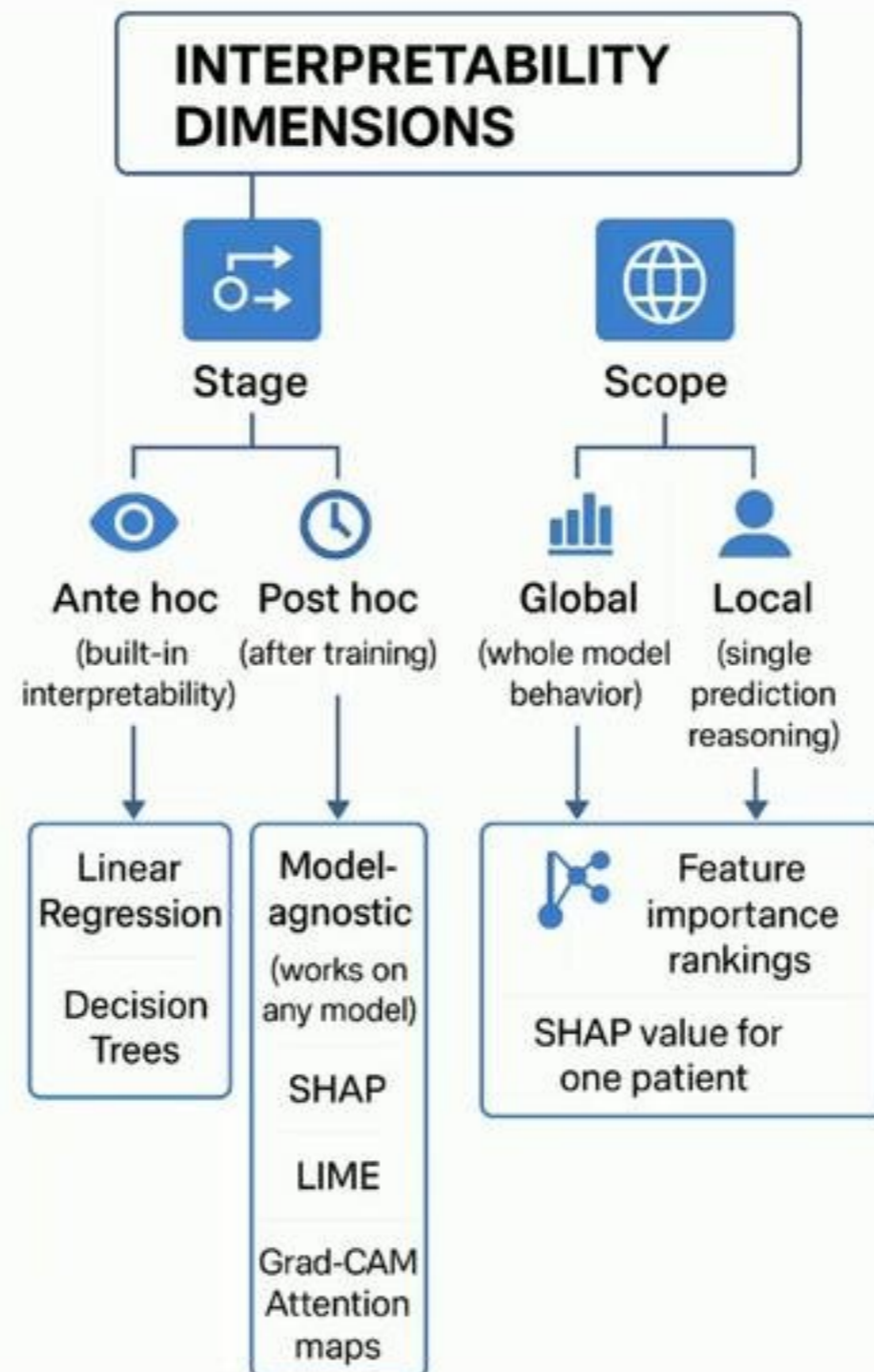
Must match the complexity & capability of the consumer

## The HealthXAI Persona Continuum

Must match the domain knowledge of the consumers

model centric aix		user centric aix	
Data Scientist		Clinical Researcher	
<b>Domain Knowledge:</b> <ul style="list-style-type: none"> <li>Speaks "AI, ML, Stats"</li> <li>Attention schemes</li> <li>Shapelets</li> <li>Model agnostic Models</li> </ul>		<b>Domain Knowledge:</b> <ul style="list-style-type: none"> <li>Clinical background</li> <li>Medical Informatics,</li> <li>Speaks AI, ML, Stats but "applied"</li> <li>Speaks/understand medical concepts</li> </ul>	
<b>Main AIX Focus:</b> <ul style="list-style-type: none"> <li>Model performance</li> <li>Model transparency</li> <li>Model and instance level explanations</li> </ul>		<b>Main AIX Focus:</b> <ul style="list-style-type: none"> <li>Model performance</li> <li>Model transparency</li> <li>Model and instance level explanations</li> <li>KPI analysis</li> </ul>	
<b>Role:</b> Direct conversations with clinician researcher		<b>Role:</b> Bidirectional conversations with both data scientist and clinicians	

Roles of personas for Disease Progression Modeling		
<b>Domain Knowledge:</b> <ul style="list-style-type: none"> <li>Apply the best suitable ML/AI technique (e.g., HMM/ RNN) for DPM</li> <li>Extend the algorithm to suit the purpose of generating interpretable DPM</li> <li>Find optimal number of stages with feedback from clinical researcher</li> </ul>	<b>Domain Knowledge:</b> <ul style="list-style-type: none"> <li>Speaks/understands the mental model of the disease progression mechanism</li> <li>Defines the inputs and outputs of overall DPM</li> </ul>	<b>Domain Knowledge:</b> <ul style="list-style-type: none"> <li>Provides clinical background about the disease mechanism</li> <li>Defines the overall goal of the DPM which can mimic the clinical progression of the disease</li> </ul>
<b>Main AIX Focus:</b> <ul style="list-style-type: none"> <li>Find explanations for each of the states of DPM</li> <li>Find instance level explanations for DPM</li> <li>Assess model performance and transparency of DPM</li> </ul>	<b>Main AIX Focus:</b> <ul style="list-style-type: none"> <li>Analyses performance and transparency of DPM</li> <li>Provides feedback about whether the stages of DPM correspond to the mental model</li> <li>Analyses the model level explanations for generating actionable insights</li> <li>Validates the Instance level explanations</li> <li>KPI analysis of the explanations</li> </ul>	<b>Main AIX Focus:</b> <ul style="list-style-type: none"> <li>Analyses the usefulness of explanation for clinical decision making</li> <li>Instance level explanations for designing patient's interventions</li> <li>Relates DPM to generate evidence based medicine</li> </ul>



Must match the complexity & capability of the consumer

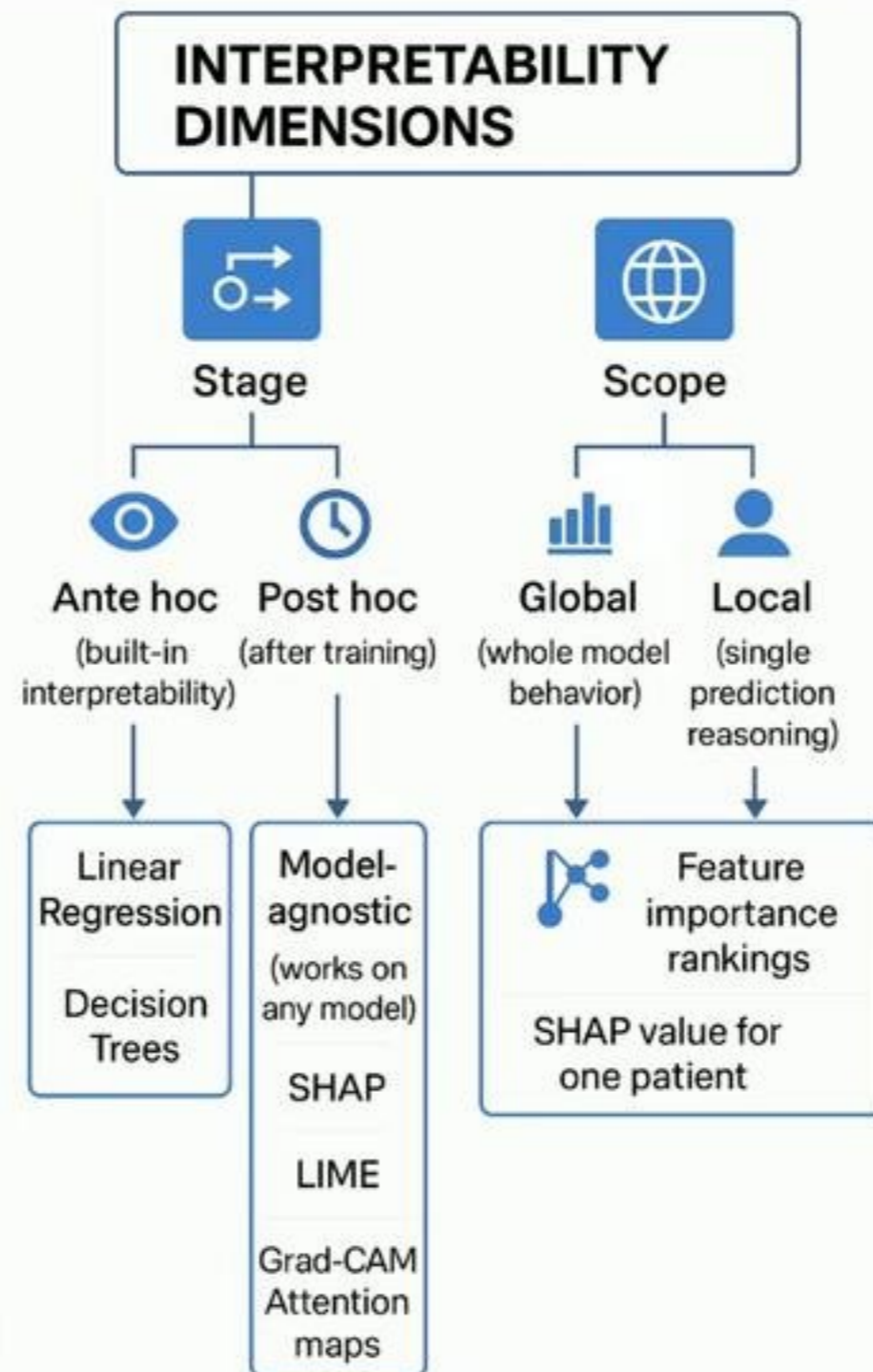
## The HealthXAI Persona Continuum

Must match the domain knowledge of the consumers

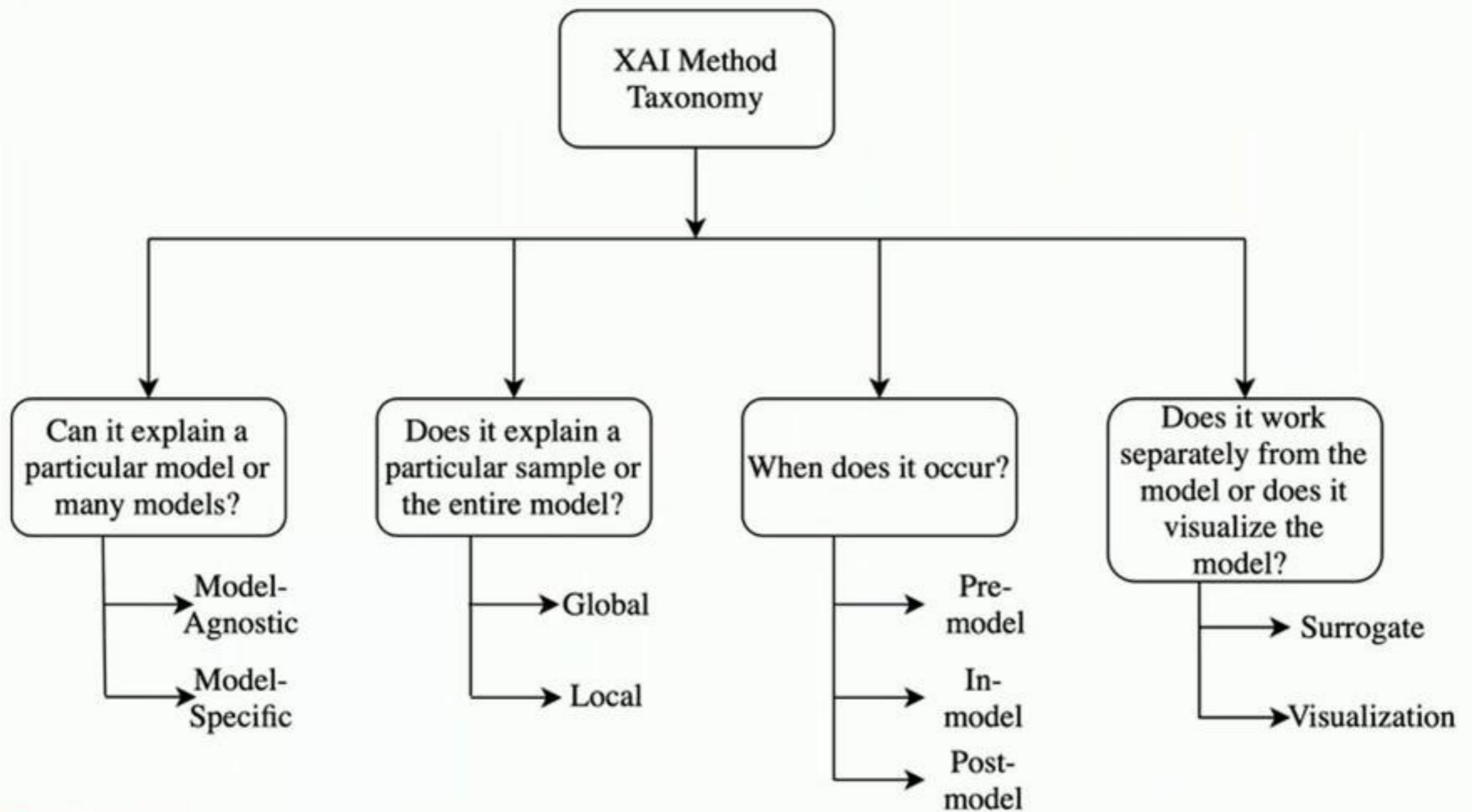
model centric aix ←		→ user centric aix	
Data Scientist		Clinical Researcher	Clinician
<b>Domain Knowledge:</b> <ul style="list-style-type: none"> <li>Speaks "AI, ML, Stats"</li> <li>Attention schemes</li> <li>Shapelets</li> <li>Model agnostic Models</li> </ul>		<b>Domain Knowledge:</b> <ul style="list-style-type: none"> <li>Clinical background</li> <li>Medical Informatics,</li> <li>Speaks AI, ML, Stats but "applied"</li> <li>Speaks/understand medical concepts</li> </ul>	
<b>Main AIX Focus:</b> <ul style="list-style-type: none"> <li>Model performance</li> <li>Model transparency</li> <li>Model and instance level explanations</li> </ul>		<b>Main AIX Focus:</b> <ul style="list-style-type: none"> <li>Model performance</li> <li>Model transparency</li> <li>Model and instance level explanations</li> <li>KPI analysis</li> </ul>	
<b>Role:</b> Direct conversations with clinician researcher		<b>Role:</b> Bidirectional conversations with both data scientist and clinicians	
<b>Domain Knowledge:</b> <ul style="list-style-type: none"> <li>Clinical expert, speaks in medical terms</li> <li>Understand medical informatics</li> <li>May have a statistics background</li> </ul>		<b>Domain Knowledge:</b> <ul style="list-style-type: none"> <li>Clinical expert, speaks in medical terms</li> <li>Understand medical informatics</li> <li>May have a statistics background</li> </ul>	
<b>Main AIX Focus:</b> <ul style="list-style-type: none"> <li>Patient centric focus</li> <li>Instance level explanations</li> </ul>		<b>Main AIX Focus:</b> <ul style="list-style-type: none"> <li>Patient centric focus</li> <li>Instance level explanations</li> </ul>	
<b>Role:</b> Direct conversations with clinician researcher		<b>Role:</b> Direct conversations with clinician researcher	

## Roles of personas for Disease Progression Modeling

<b>Domain Knowledge:</b> <ul style="list-style-type: none"> <li>Apply the best suitable ML/AI technique (e.g., HMM/ RNN) for DPM</li> <li>Extend the algorithm to suit the purpose of generating interpretable DPM</li> <li>Find optimal number of stages with feedback from clinical researcher</li> </ul>	<b>Domain Knowledge:</b> <ul style="list-style-type: none"> <li>Speaks/understands the mental model of the disease progression mechanism</li> <li>Defines the inputs and outputs of overall DPM</li> </ul>	<b>Domain Knowledge:</b> <ul style="list-style-type: none"> <li>Provides clinical background about the disease mechanism</li> <li>Defines the overall goal of the DPM which can mimic the clinical progression of the disease</li> </ul>
<b>Main AIX Focus:</b> <ul style="list-style-type: none"> <li>Find explanations for each of the states of DPM</li> <li>Find instance level explanations for DPM</li> <li>Assess model performance and transparency of DPM</li> </ul>	<b>Main AIX Focus:</b> <ul style="list-style-type: none"> <li>Analyses performance and transparency of DPM</li> <li>Provides feedback about whether the stages of DPM correspond to the mental model</li> <li>Analyses the model level explanations for generating actionable insights</li> <li>Validates the Instance level explanations</li> <li>KPI analysis of the explanations</li> </ul>	<b>Main AIX Focus:</b> <ul style="list-style-type: none"> <li>Analyses the usefulness of explanation for clinical decision making</li> <li>Instance level explanations for designing patient's interventions</li> <li>Relates DPM to generate evidence based medicine</li> </ul>



Type	Method
Local Explanations	<b>Feature Importance</b> Explainer assigns to each feature an importance value which represents how much that particular feature was important for the prediction under analysis
	<b>Rule Based</b> Explicitly state the decision support system's decision boundary between the given and the contrasting advice, which can be viewed as "if... then..." statements
	<b>Saliency Maps</b> Generally used with image or video processing applications and is supposed to show what parts are most important to a network's decisions
	<b>Prototypes Based</b> A prototype is an object representing a set of similar records that the user can easily view, understand and appreciate the similarity to other validation samples
	<b>Counterfactuals</b> Explanation that provides a link between what could have happened had the input to a model been changed in a particular way
Global Explanations	<b>Collection of Local Explanations</b> Pick subset of k local explanations to constitute the global explanation after generating a local explanation for every data instance using one of the above approaches
	<b>Representation Based</b> Derive model understanding by analyzing intermediate representations (DNN) and determine model's reliance on 'concepts' that are semantically meaningful to humans
	<b>Model Distillation</b> Leverage model distillation to learn feature shapes that describe the relationship between input features and model predictions
	<b>Summaries of Counterfactuals</b> Construct global counterfactual explanations which provide an interpretable and accurate summary of recourses for the entire population



# Tabular Data

## Tabular

### Feature Importance Based Explanations



### Rule Based Explanations

if hemiplegia and age > 60 then stroke risk 58.9% (53.8%-63.8%)  
 else if cerebrovascular disorder then stroke risk 47.8% (44.8%-50.7%)  
 else if transient ischaemic attack then stroke risk 23.8% (19.5%-28.4%)  
 else if occlusion and stenosis of carotid artery without infarction then stroke risk 15.8% (12.2%-19.6%)  
 else if altered state of consciousness and age > 60 then stroke risk 16.0% (12.2%-20.2%)  
 else if age ≤ 70 then stroke risk 4.6% (3.9%-5.4%)  
 else stroke risk 8.7% (7.9%-9.6%)

Letham et al., 2015

### Prototype Based Explanations

Top 5 instances driving the prediction

Day	Outlook	Temp	Humid	Wind	Play?
1	Sunny	Cool	Normal	Weak	Yes
2	Rain	Mild	Normal	Weak	Yes
3	Sunny	Mild	Normal	Strong	Yes
4	Overcast	Mild	High	Strong	Yes
5	Overcast	Hot	Normal	Weak	Yes

Prediction: Play

### Counterfactual Explanations



Recourse: Your loan will be approved if you increase your income by \$25000 and close 3 accounts

# Text Data

Text

## Saliency Map Visualization

### Simple Gradients Visualization

See saliency map interpretations generated by visualizing the gradient.

Saliency Map:

[CLS] The [MASK] rushed to the emergency room to see her patient. [SEP]

### Mask 1 Predictions:

47.1% nurse  
16.4% woman  
10.0% doctor  
3.4% mother  
3.0% girl

Wallace et al., 2019

## Input Reduction

A puzzling man named NLP Cool went to buy some organic fruit at Grandpa Joe's in downtown Deep Learning

Reduced input for NLP Cool named NLP Cool  
Reduced input for Grandpa Joe's at Grandpa Joe's  
Reduced input for Deep Learning in downtown Deep Learning

Wallace et al., 2019

**SNE**  
Premise: Well dressed man and woman dancing in the street.  
Original: Two man is dancing on the street dancing.  
Reduced: Contradiction.  
Answer: 0.977 → 0.706.

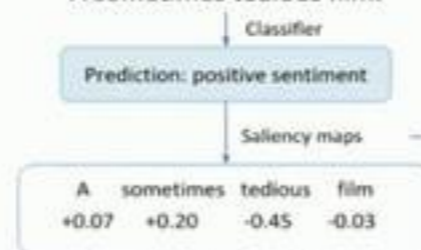
**VQA**  
Original: What color is the flower?  
Reduced: flower?  
Answer: yellow  
Confidence: 0.827 → 0.819



Feng et al., 2018

## Prototype Based Explanations

A sometimes tedious film.



Salient tokens in the input

Influence functions

Credulous. positive +10.32  
An admittedly middling film. positive +10.09  
A simplistic narrative. positive +9.58  
⋮  
Tedious Norwegian offering which somehow snagged an oscar nomination. negative -9.64  
Visually flashy but narratively opaque. negative -11.01  
Full of cheesy dialogue. negative -12.78

Influential examples in the training corpus

Han et al., 2020

# Image Data

Image

Saliency Map Visualization



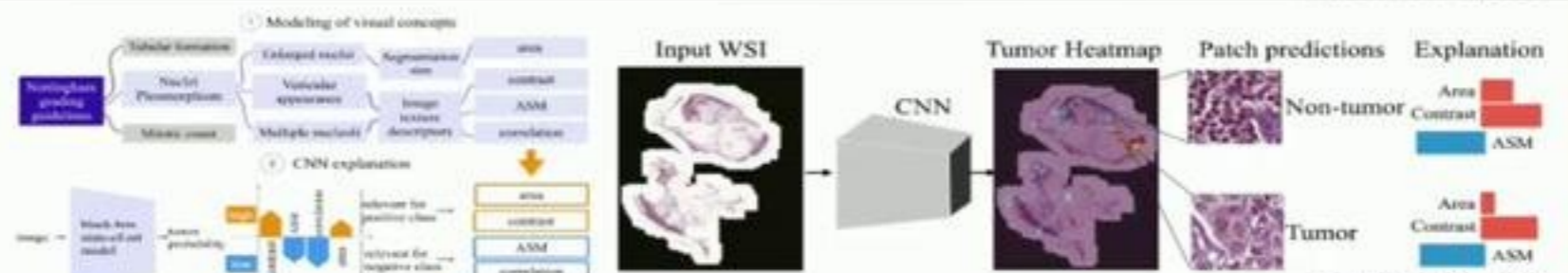
Simonyan et al., 2013

Shapley Value Importance



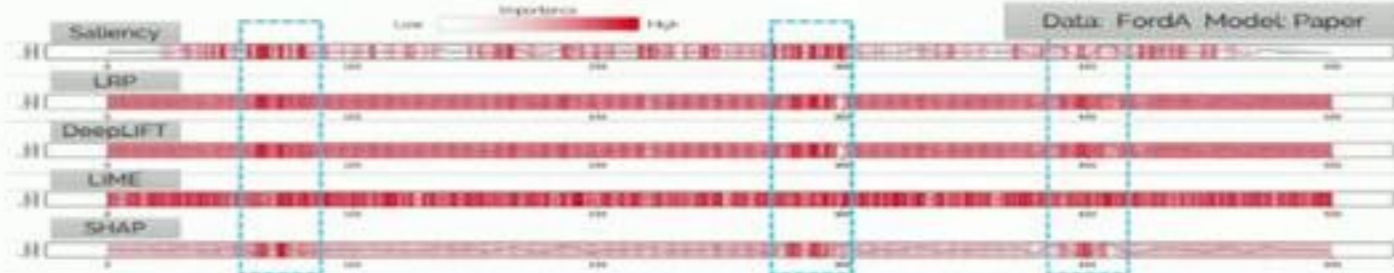
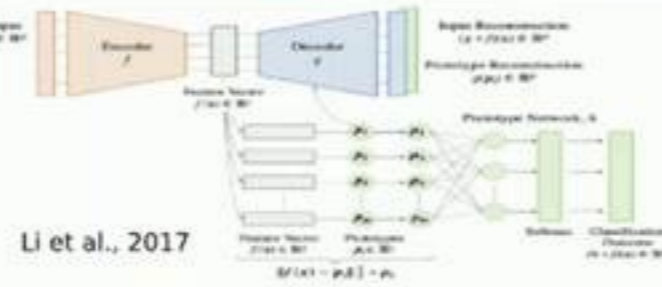
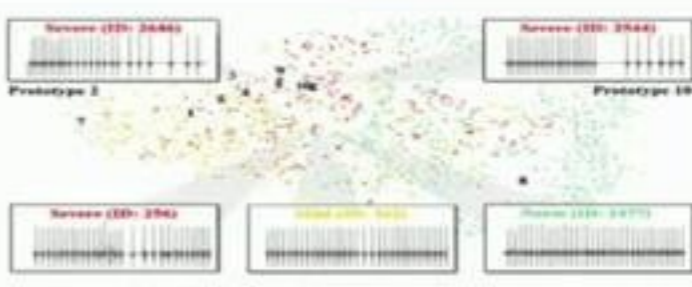
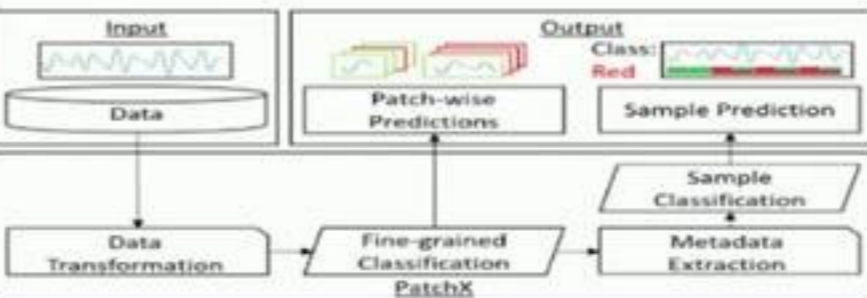
Ghorbani et al., 2020

Concept Attribution



Graziani et al., 2020

# Time Series Data

Time Series	Relevance Heatmaps	 <p>Saliency LRP DeepLIFT LIME SHAP</p> <p>Low High Importance</p> <p>Data: FordA Model: Paper</p> <p>Schlegel et al., 2019</p>
	Prototype Based Explanations	 <p>Li et al., 2017</p>  <p>Gee et al., 2019</p>
	Patch Based Classification	 <p>Input Data Data Transformation Patch-wise Predictions Sample Prediction Sample Classification Metadata Extraction Output Class: Red</p> <p>(a) No anomaly patches (b) Anomaly patches</p> <p>Mercier et al, 2021</p>

# Going to detail of a few techniques

# Example 1: Model Dev

Chrome File Edit View History Bookmarks Profiles Tab Window Help

GoTo Meeting

app.goto.com/meeting/333782365

New Chrome available

All Bookmarks

Conference Support, PATRIC... Everyone

Conference Support

Rayan Harari

PG

You're sharing your screen

PG

PATRICIA E SORTILLON G

Record React Mic Camera Share Leave Captions Pop out

26

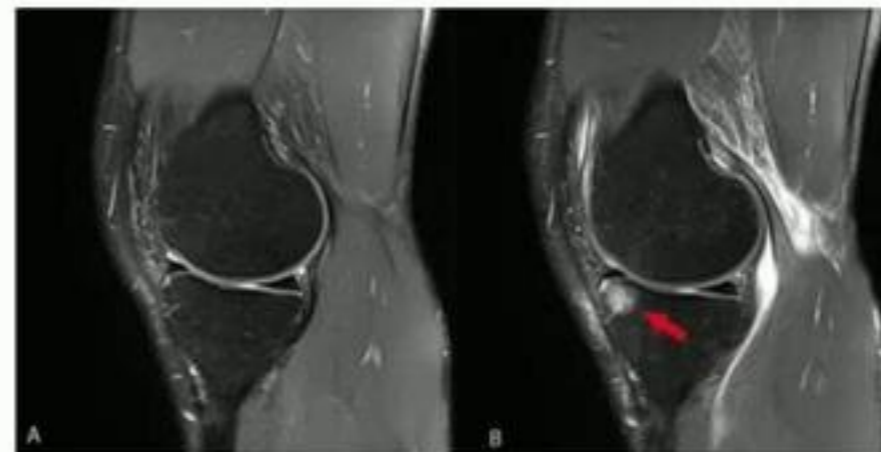
# Knee OA Progression

## Osteoarthritis (OA) Burden

- OA is a major cause of disability worldwide (4th leading cause).
- Significant unmet need for disease-modifying therapies (**DMOADS**)
- Radiographic progression is slow and heterogeneous, complicating drug trials.

## MRI

- X-ray measures (joint space width) are standard but not always sensitive enough.
- MRI provides detailed anatomical information (cartilage, meniscal tears, effusions).



## Data Inputs for Modeling

Measurement

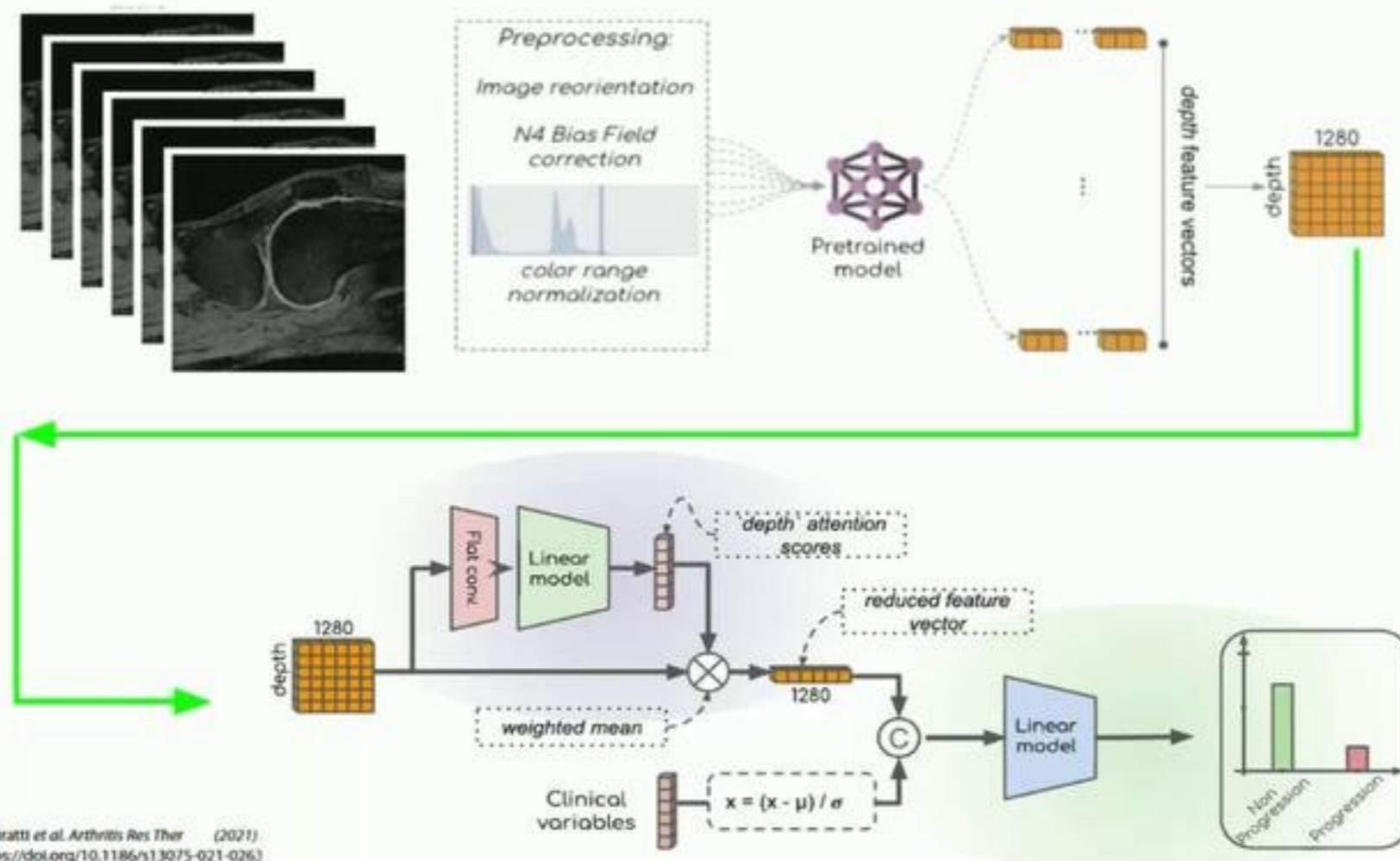
Other Clinical Data  
(Scores)+ Demog

Images (for  
CNN)

WOMAC	Medi	SBF	Cart Vol (BL)	Cart Vol (D4mo)	MF (BL)	MT (BL)	LT (BL)	LF (BL)
2	1	27.4	732.5	630.8	206.68	77.59	145.3	256.83
0	0	22.9	479.3	391.9	211.13	119.74	141.56	229.97
0	0	34.7	1028	1000.7	0	2.05	49.05	28.35
2	0	27.3	385.3	281.3	97.17	83	0	23.87
1	0	25.6	683.9	290.1	49.43	56.14	26.16	0
0	1	30.7	577.2	309.1	137.46	29.66	108.92	278.84
2	1	32.7	737.6	856.3	43.83	61.74	16.04	0
4	0	38.2	496.9	482.5	117.88	223.07	11	84.86
2	0	33.4	545.6	587.5	35.81	44.95	0	42.9
0	1	32.9	602	576.2	34.32	0	107.06	9.7
8	0	34.8	583	472.8	36.56	44.76	57.45	191.36
0	1	29.4	395.6	327.2	77.22	140.07	35.62	150.33
8	1	31.2	619.5	502	119.06	90.09	211.69	314.84
2	0	24.2	891.5	551.7	70.69	8.95	68.26	79.85
0	0	30.2	425.3	378.9	40.66	32.45	0	80.57
0	1	24.6	352.9	304.8	66.77	28.16	29.84	59.68
8	0	30.8	953.7	825.1	317.08	56.51	306.07	74.23
8	0	35.3	279.2	289.1	1356.24	183.18	423.2	336.29
1	0	31.4	364.2	399	168.05	81.51	16.79	85.24
5	0	24.7	864.6	846.3	41.41	194.54	125.9	212.07
0	0	27.8	571.5	596.3	374.52	290	73.3	129.83
1	0	30.2	422.6	398.5	7.83	36	95.5	62.46
0	0	30.8	994.9	1006.9	0	12.87	147.16	274.55



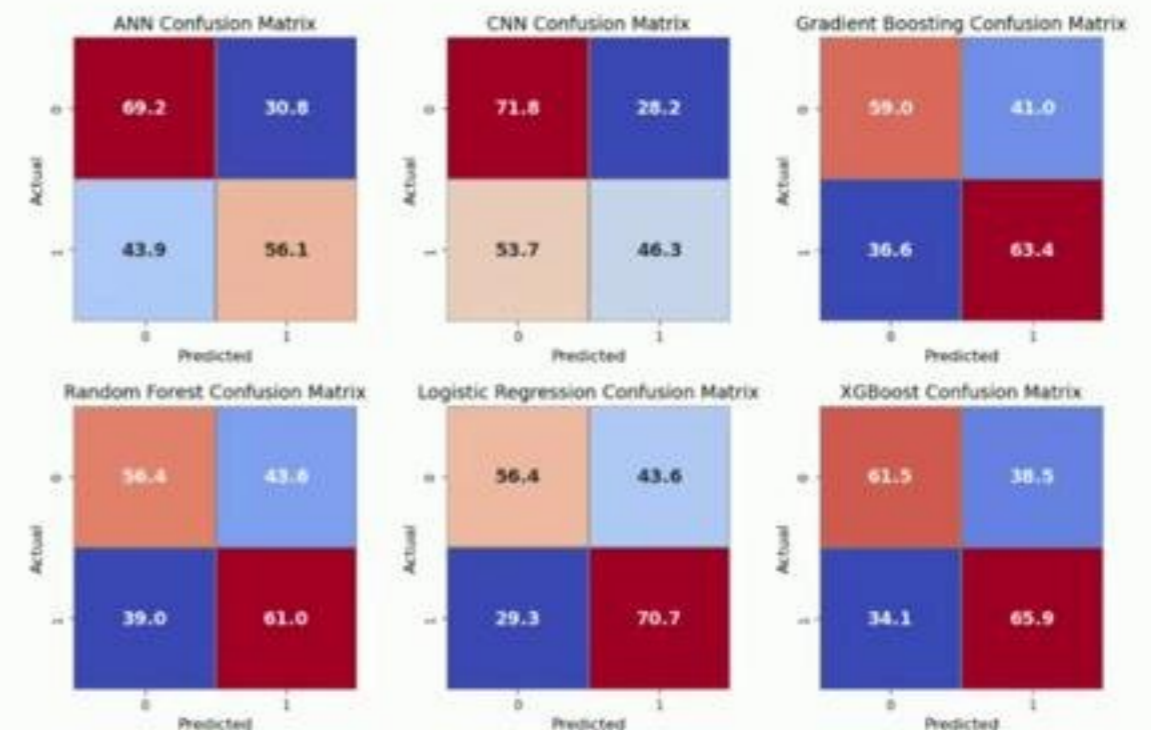
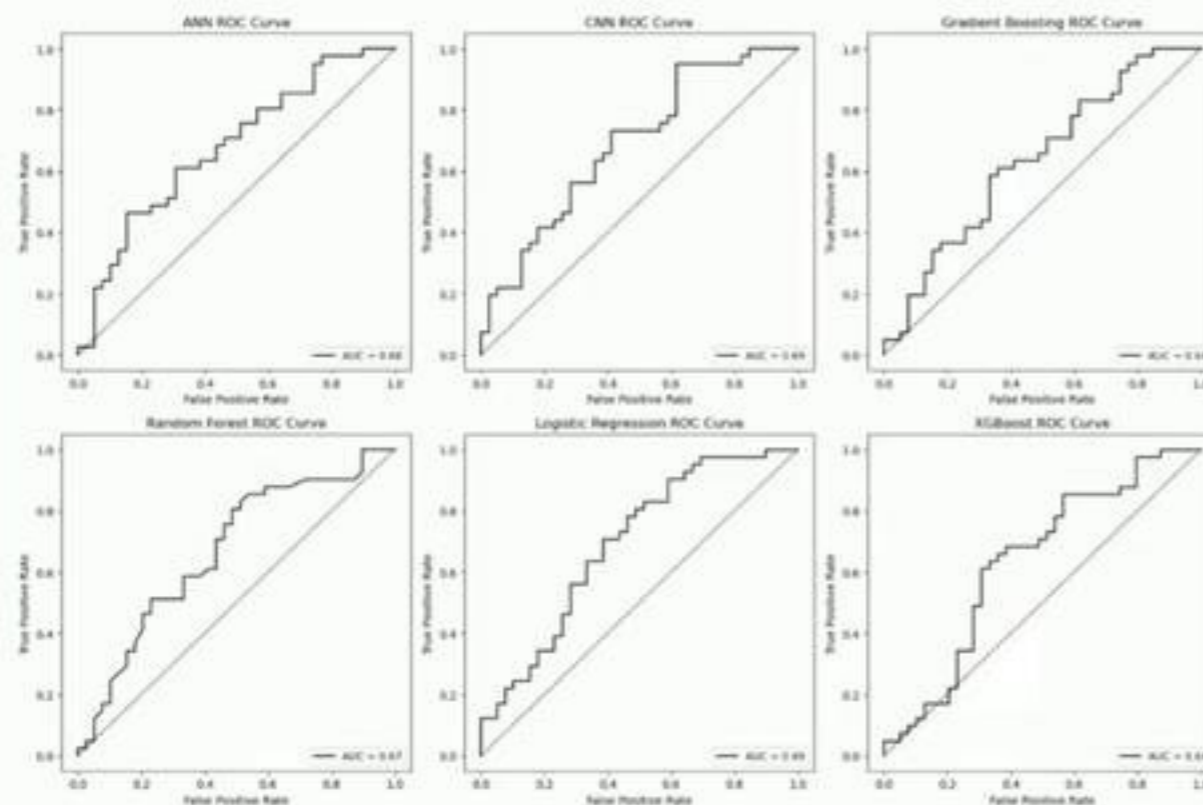
# A Multimodal Approach



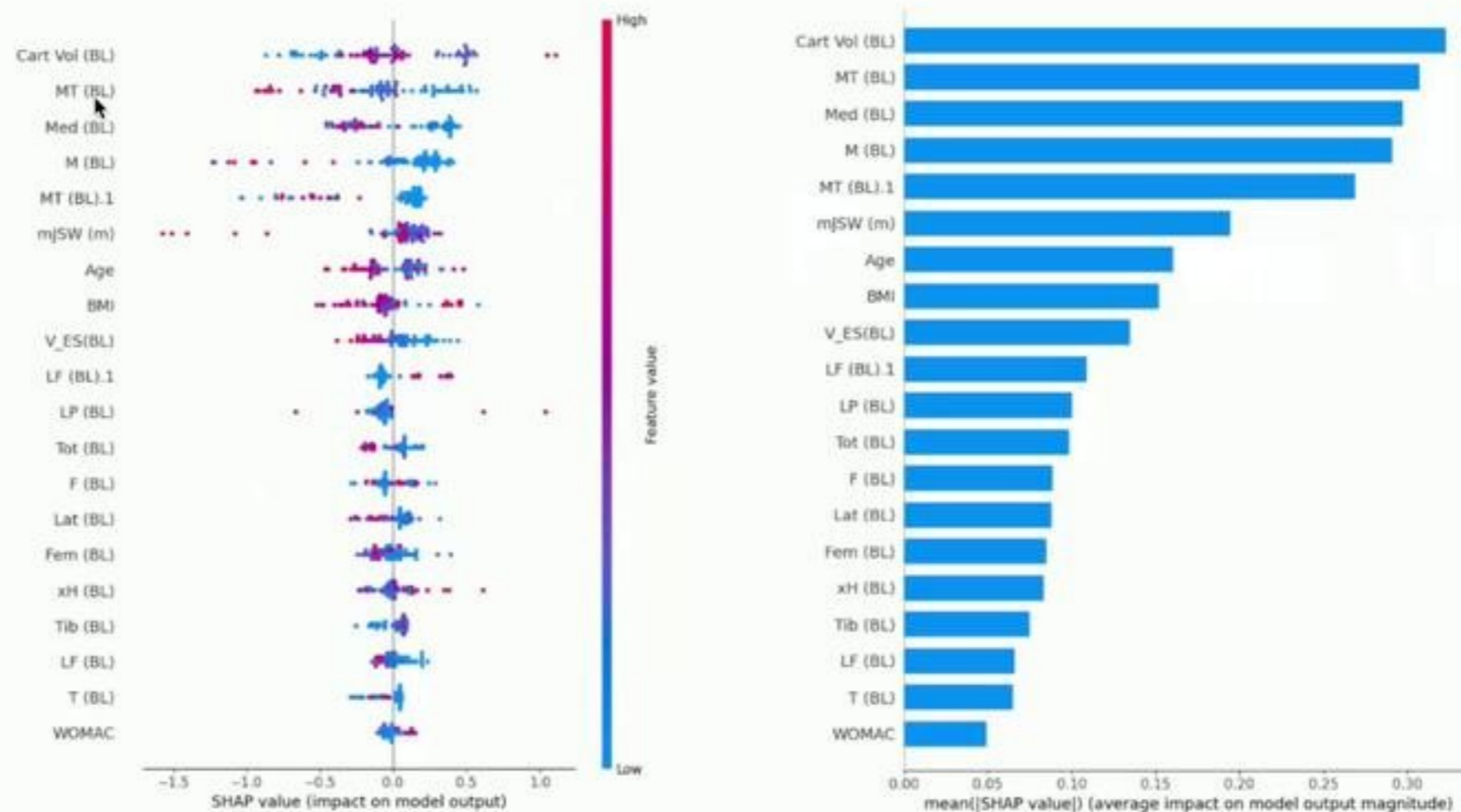
Schiratti et al. Arthritis Res Ther (2021)  
<https://doi.org/10.1186/s13075-021-0263>

- **Feature Extractor** (EfficientNet-B0)
  - Pre-trained on ImageNet, outputting 1280-d embeddings per slice.
- **Attention Sub-Model**
  - Learns slice "importance" weights (particularly relevant slices for cartilage changes).
- **Multilayer Perceptron**
  - Aggregates weighted MRI features + clinical data → final classifier (progressor vs. non-progressor).
- Hyperparameter tuning (learning rate, batch size) with cross-validation.

# ML Models Performance on Measurements/Clinical Data



# ML Models Explainability (xAI)

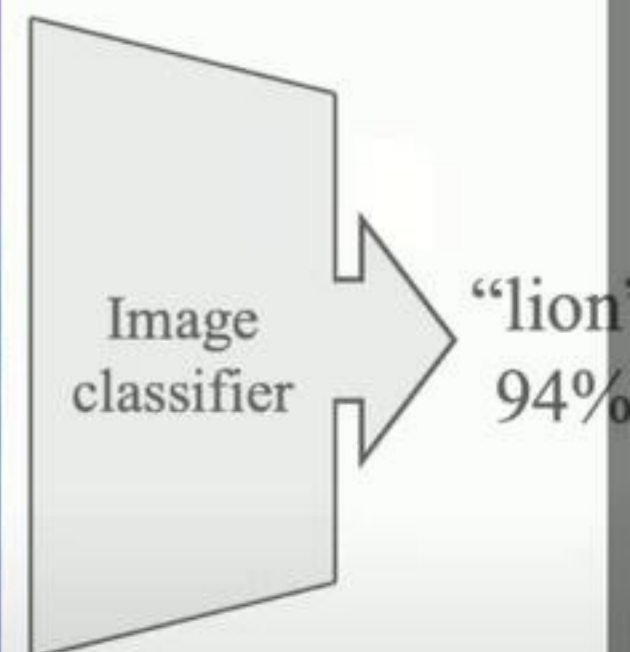


# Example 1: Alg Dev

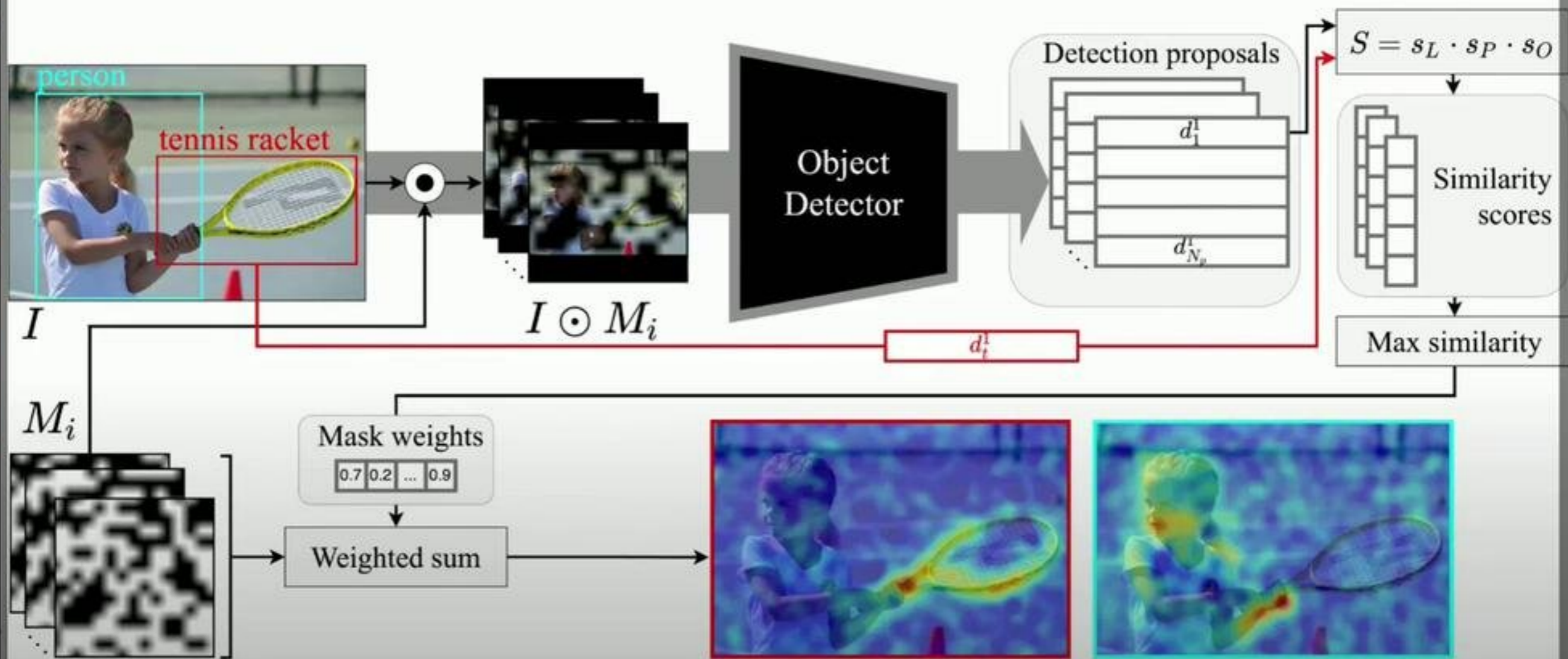
# Saliency maps

Saliency maps have become a popular tool for analyzing and explaining neural network models.

These heatmaps show which regions of the input are more important for the model's prediction.

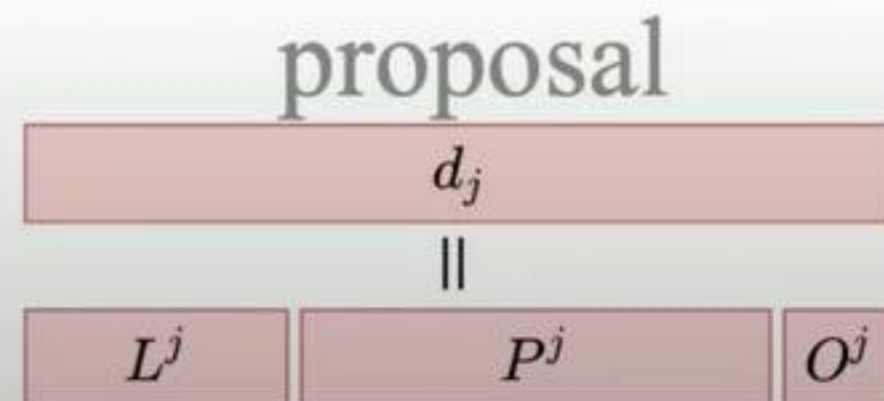
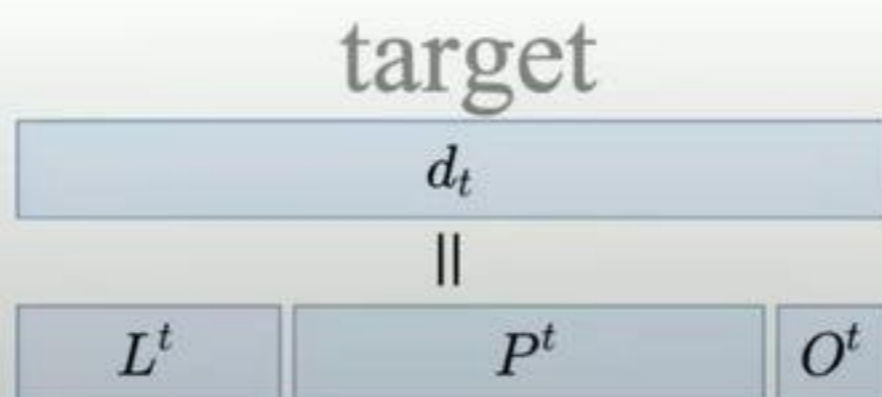


# Saliency maps



# Measuring the effect of input masking

- Localization = bounding box coordinates.
- Classification = class probabilities.
- Objectness =  $[0, 1]$  scalar.

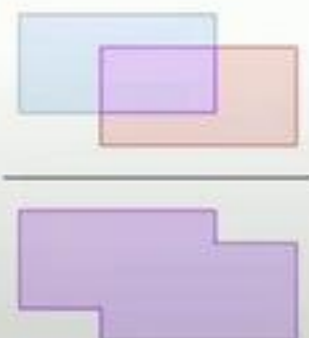


# Measuring the effect of input masking

Similarity score is computed using the similarities between the three individual components.

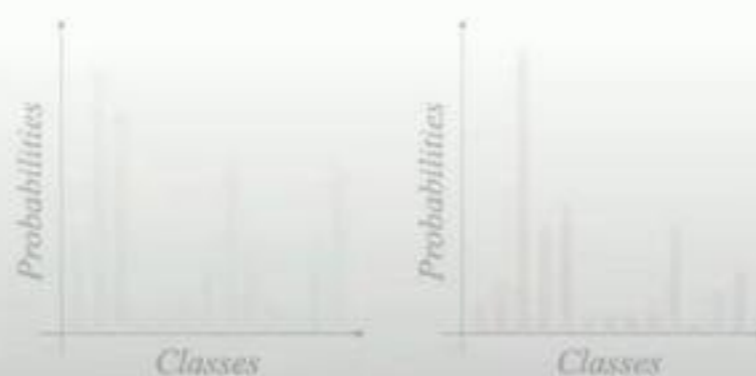
$L^t$	$P^t$	$O^t$
$L^j$	$P^j$	$O^j$

*Intersection over Union*



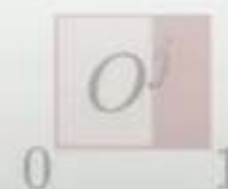
$$s_L(d_t, d_j) = \text{IoU}(L^t, L^j)$$

*Cosine similarity*



$$s_P(d_t, d_j) = \frac{P^t \cdot P^j}{\|P^t\| \|P^j\|}$$

*Objectness score (Optional)*



$$s_O(d_t, d_j) = O^j$$

# Measuring the effect of input masking

Similarity score is computed using the similarities between the three individual components.

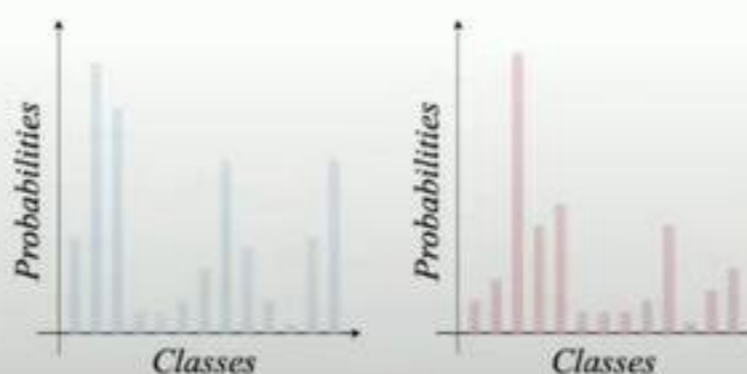
$L^t$	$P^t$	$O^t$
$L^j$	$P^j$	$O^j$

*Intersection over Union*



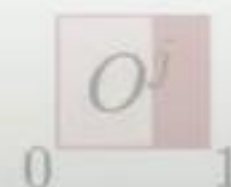
$$s_L(d_t, d_j) = \text{IoU}(L^t, L^j)$$

*Cosine similarity*



$$s_P(d_t, d_j) = \frac{P^t \cdot P^j}{\|P^t\| \|P^j\|}$$

*Objectness score (Optional)*



$$s_O(d_t, d_j) = O^j$$

# Measuring the effect of input masking

Similarity score is computed using the similarities between the three individual components.

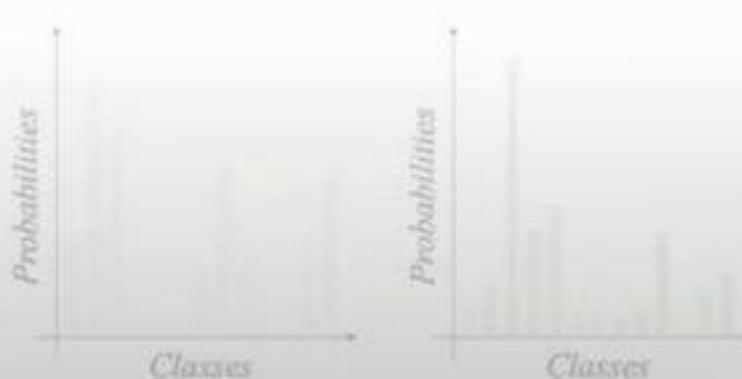
$L^t$	$P^t$	$O^t$
$L^j$	$P^j$	$O^j$

*Intersection over Union*



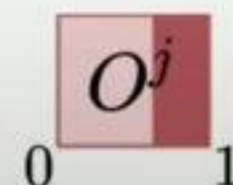
$$s_L(d_t, d_j) = \text{IoU}(L^t, L^j)$$

*Cosine similarity*



$$s_P(d_t, d_j) = \frac{P^t \cdot P^j}{\|P^t\| \|P^j\|}$$

*Objectness score (Optional)*



$$s_O(d_t, d_j) = O^j$$

# Measuring the effect of input masking

Similarity score is computed using the similarities between the three individual components.

$L^t$	$P^t$	$O^t$
$L^j$	$P^j$	$O^j$

*Intersection over Union*

*Cosine similarity*

*Objectness score (Optional)*

$$s(d_t, d_j) = s_L(d_t, d_j) \cdot s_P(d_t, d_j) \cdot s_O(d_t, d_j)$$

$$s_L(d_t, d_j) = \text{IoU}(L^t, L^j)$$

$$s_P(d_t, d_j) = \frac{P^t \cdot P^j}{\|P^t\| \|P^j\|}$$

$$s_O(d_t, d_j) = O^j$$

# Pointing Game metric

The Pointing Game metric measures how often the saliency map peak falls within the bounding box or segmentation mask of the object. When this happens a “hit” is scored. If the peak is outside the object it is counted as a “miss”.

Final Pointing Game score reports the accuracy:

$$PG = \frac{N_{\text{hits}}}{N_{\text{hits}} + N_{\text{misses}}}$$



# User's trust

Instructions

[View full instructions](#)

[View tool guide](#)

Increasing importance

Two Robots are trying to find the same object and are showing you how they made their decision through a heat map

Help us decide if Robot 1 is much better, slightly better, or about the same as Robot 2

Two robots are showing you why they found the following object within the box. Which Robot's explanation is more reasonable?

Detected Object

Robot 1 Explanation

Robot 2 Explanation

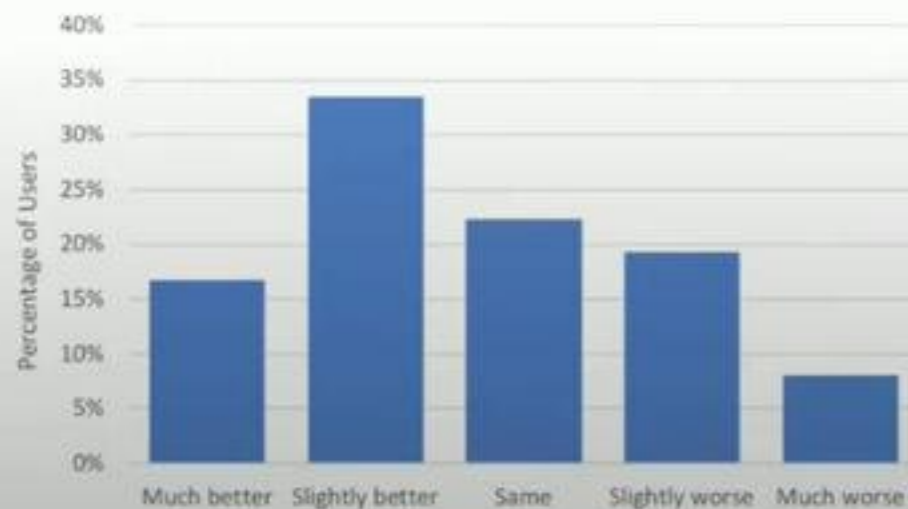
Zoom in

Zoom out

Move

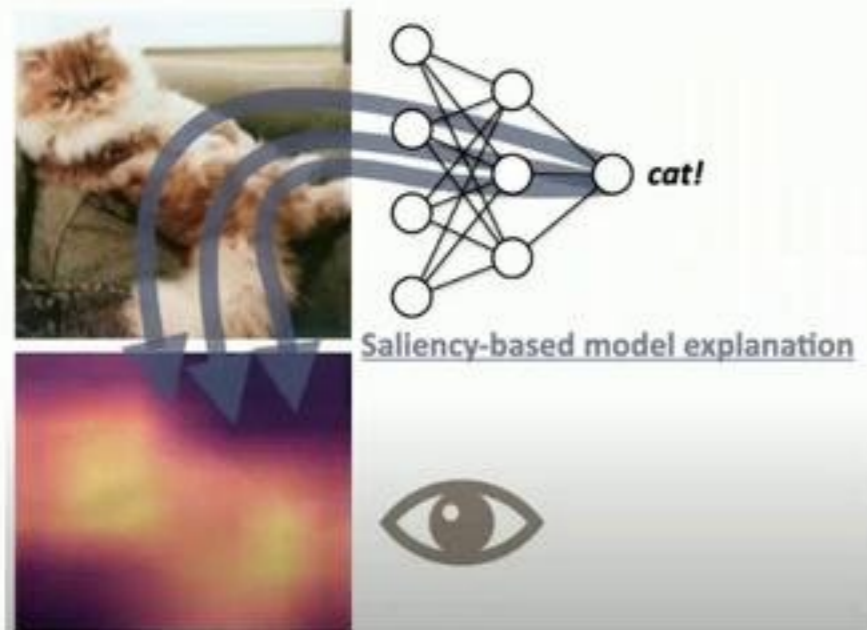
Fit Image

Submit



Substantially more users (50.2% vs 27.4%) found that explanations of a stronger model (YOLOv3 vs YOLOv3-Tiny) to be better or more trustworthy.

# Saliency Map



Two axes of visual encoding of saliency map – visual range



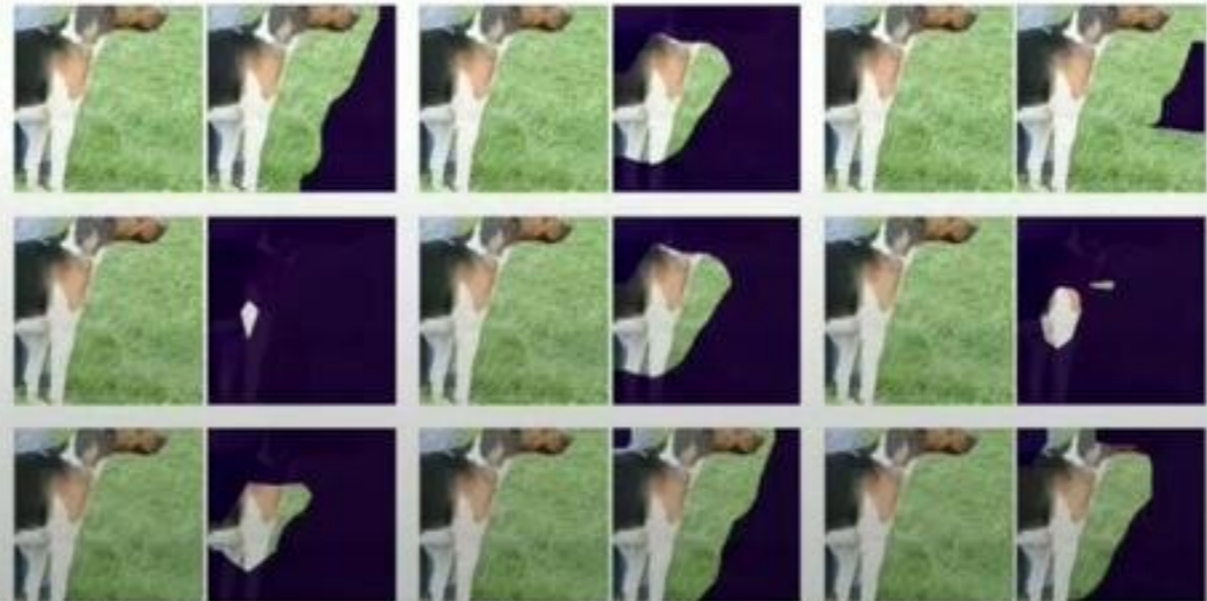
# User study interface

Please select **all** (0 to 9, you can select all or none based on your judgement)  
visualizations which *faithfully capture* the object:

Please select **all** (0 to 9, you can select all or none based on your judgement)  
visualizations which *faithfully capture* the object:

Please select **all** (0 to 9, you can select all or none based on your judgement)  
visualizations which *faithfully capture* the object:

## Beagle dog

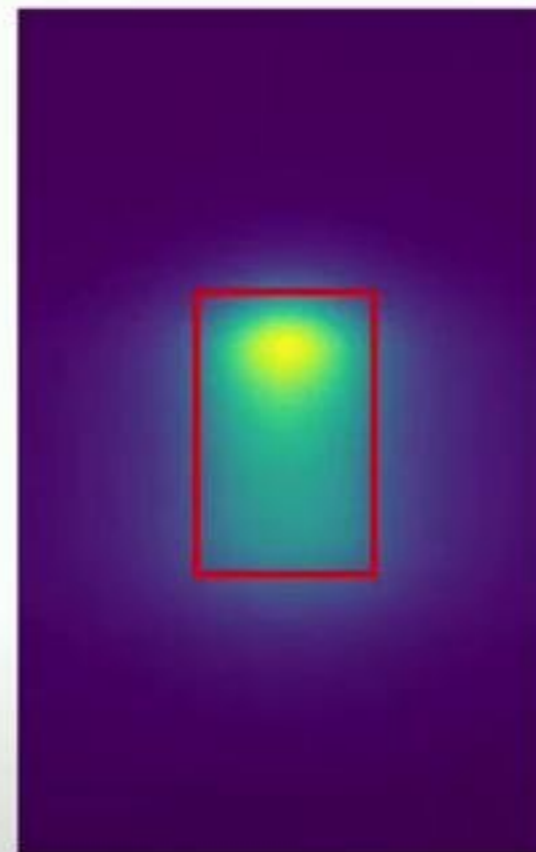


# Analyzing average saliency maps

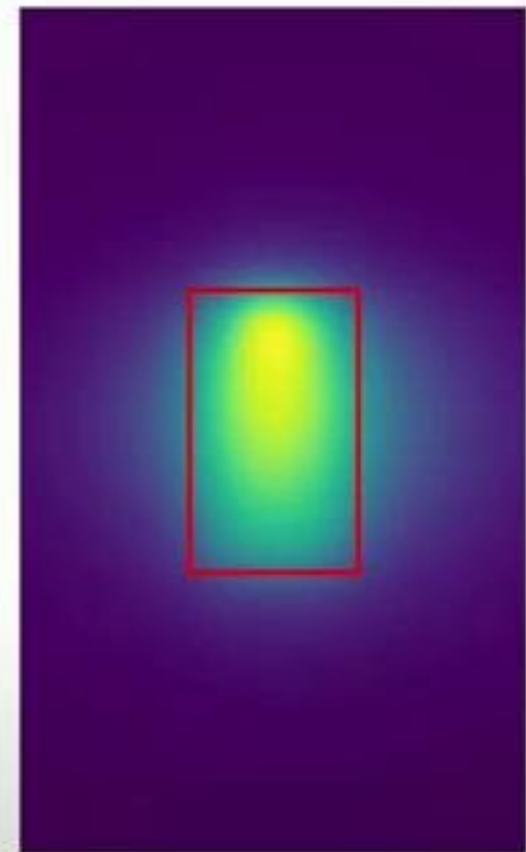
By computing average saliency maps per category we can see some common patterns in input importance.



Normalized average  
image of *person*



Average saliency  
map for YOLOv3



Average saliency  
map for Faster-RCNN

# Failure modes

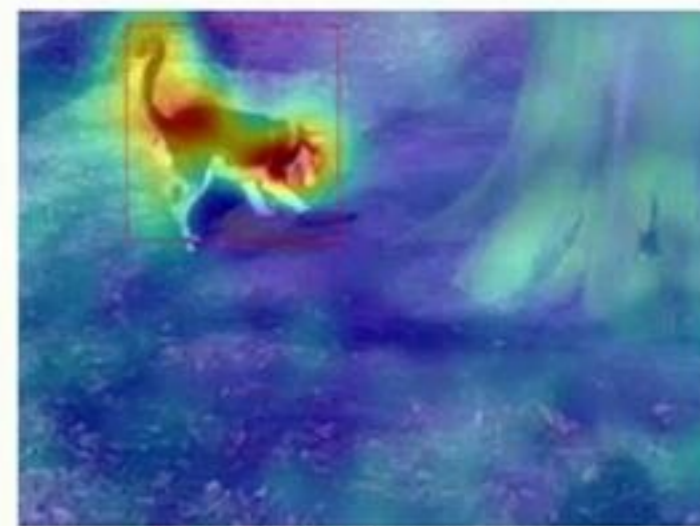
- False positive detections
- Missed detections
- **Poor classification**
- Poor localization

We can subtract normalized saliency maps to analyze what caused the classification error. Here, the dog's black fur and tail could be the cause.

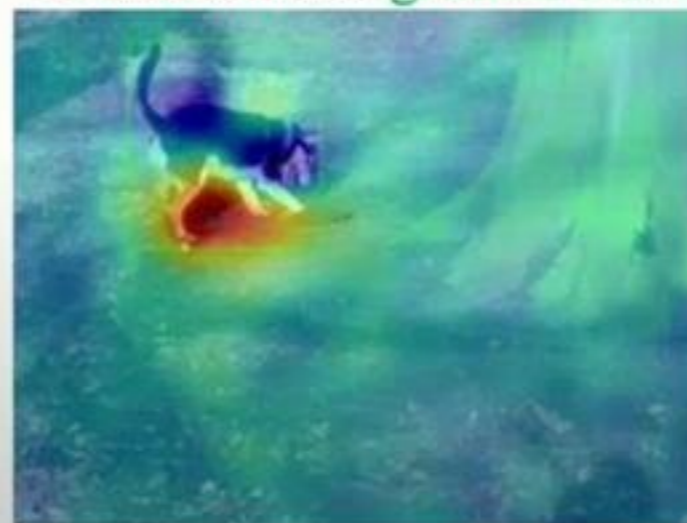
Dog predicted as Cat



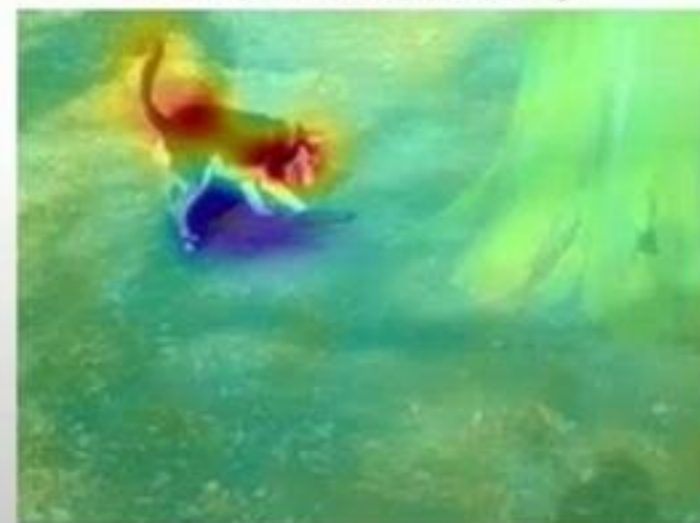
Predicted and ground truth



Predicted saliency



Ground truth saliency



$\text{Norm(P)} - \text{Norm(GT)}$

## **Example 3: A Case Study of User Understanding**



# Motivation

- To develop and test an AI decision support system with different modalities in explainability (decision tree, etc) for perfusionist during operations.

# xAI Techniques

- Decision Tree
- Counterfactual
- Case-based
- Feature Importance
- Nothing

## Counterfactual

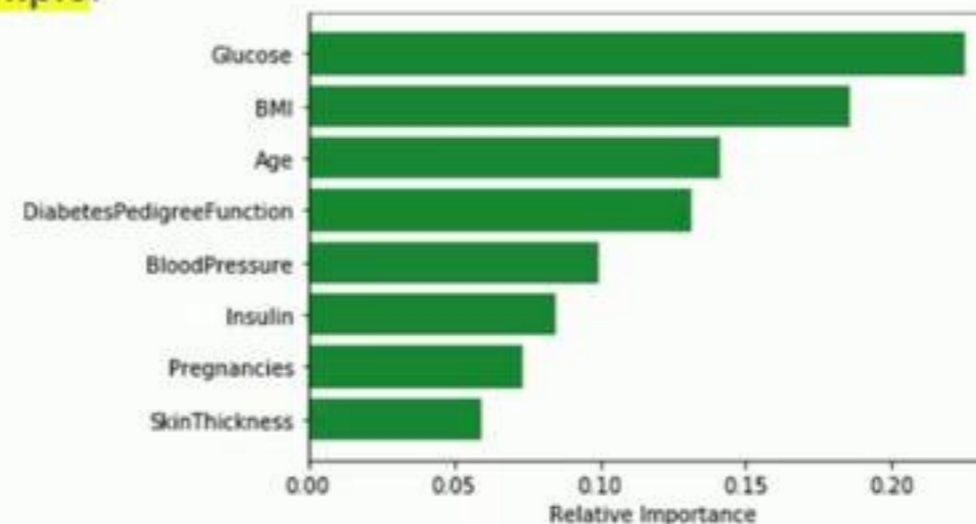
- A technique that involves generating explanations by contrasting the actual outcome with an alternative outcome that could have occurred.
- A counterfactual explanation describes a causal situation in the form: "If X had not occurred, Y would not have occurred".
- **Example:** "If I hadn't taken a sip of this hot coffee, I wouldn't have burned my tongue". Event Y is that I burned my tongue; cause X is that I had a hot coffee.

# xAI Techniques

- Decision Tree
- Counterfactual
- Case-based
- Feature Importance
- Nothing

## Feature Importance

- A technique that involves identifying the most important features or factors that contributed to a particular decision or outcome.
- **Example:**

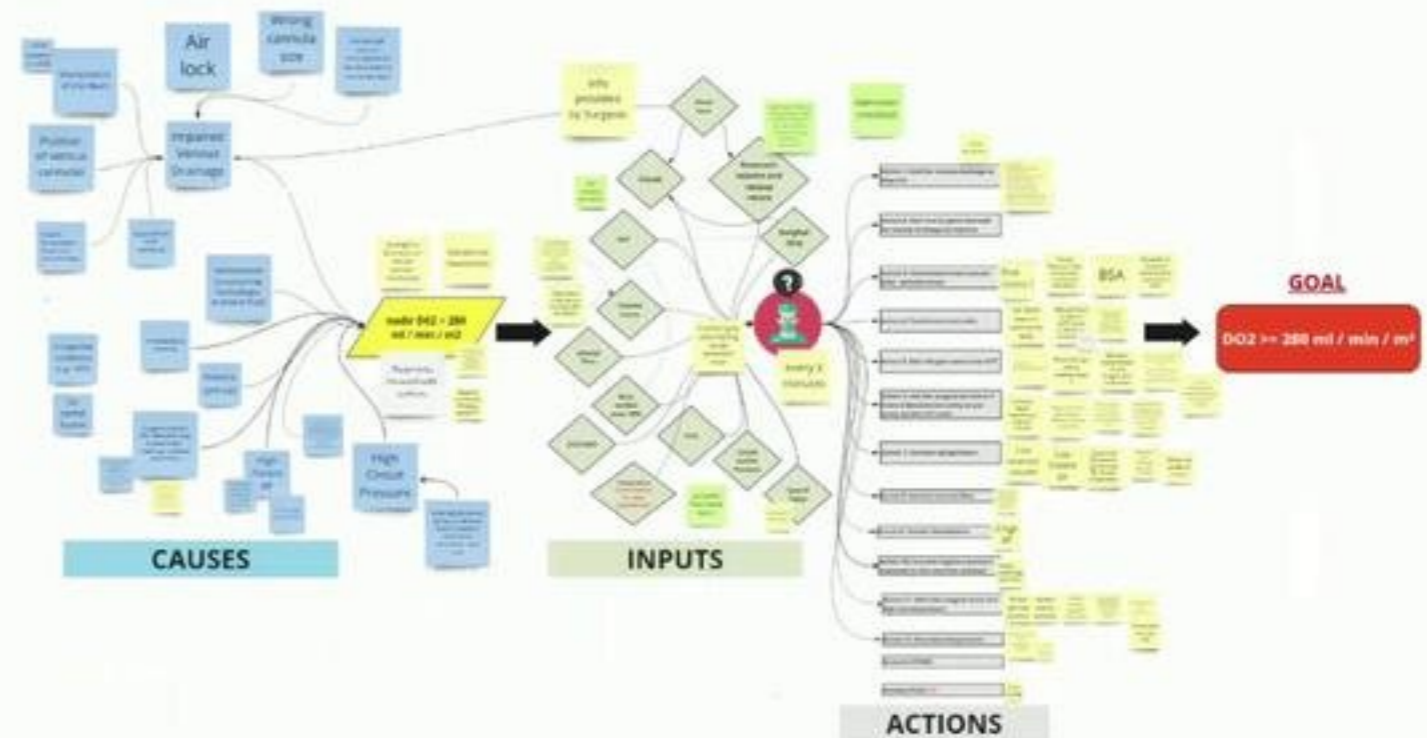


## Counterfactual

- A counterfactual explanation describes a causal situation in the form: "If X had not occurred, Y would not have occurred".
- **Example:** "If I hadn't taken a sip of this hot coffee, I wouldn't have burned my tongue". Event Y is that I burned my tongue; cause X is that I had a hot coffee.

## How can we develop an AI-based decision support system for this settings?

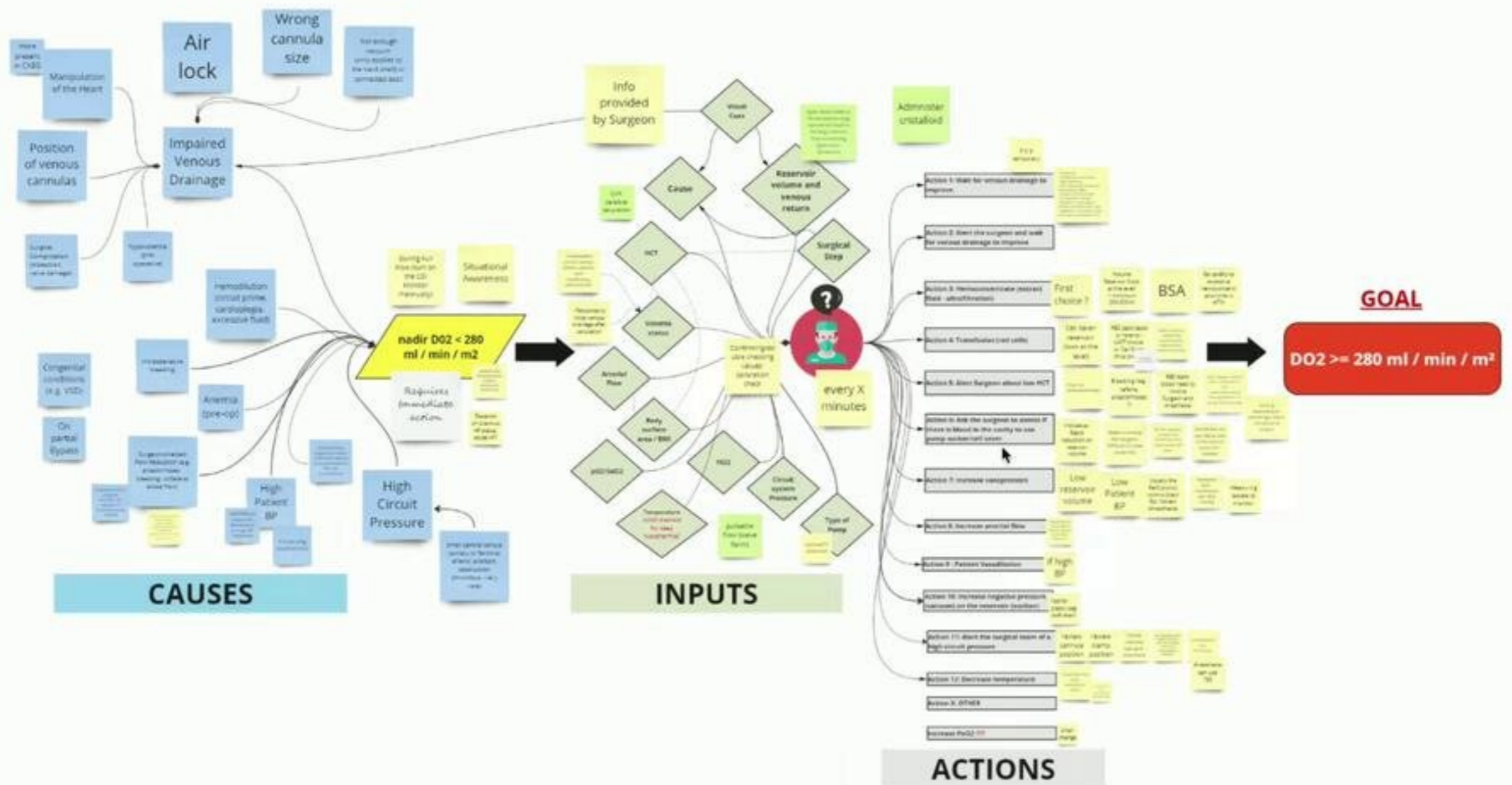
## How human naturally makes decision in high-stakes environment?

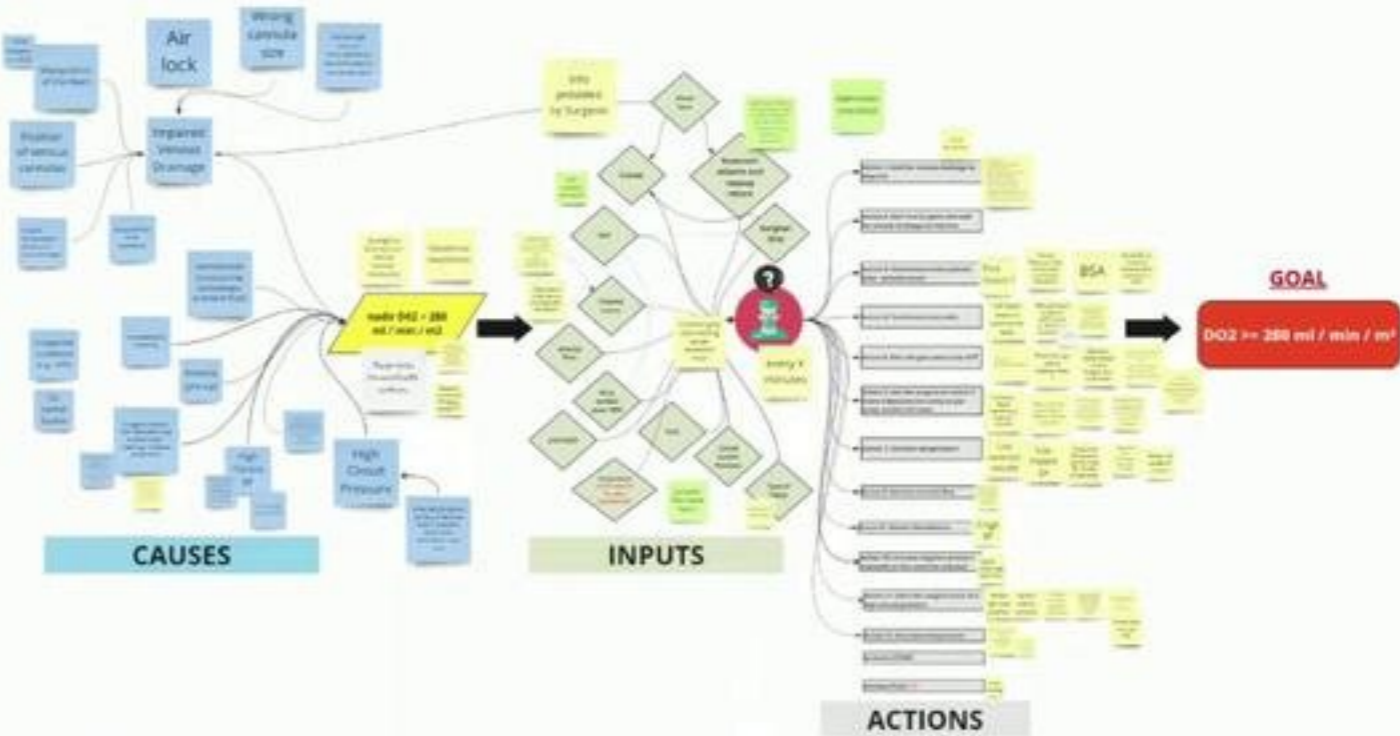


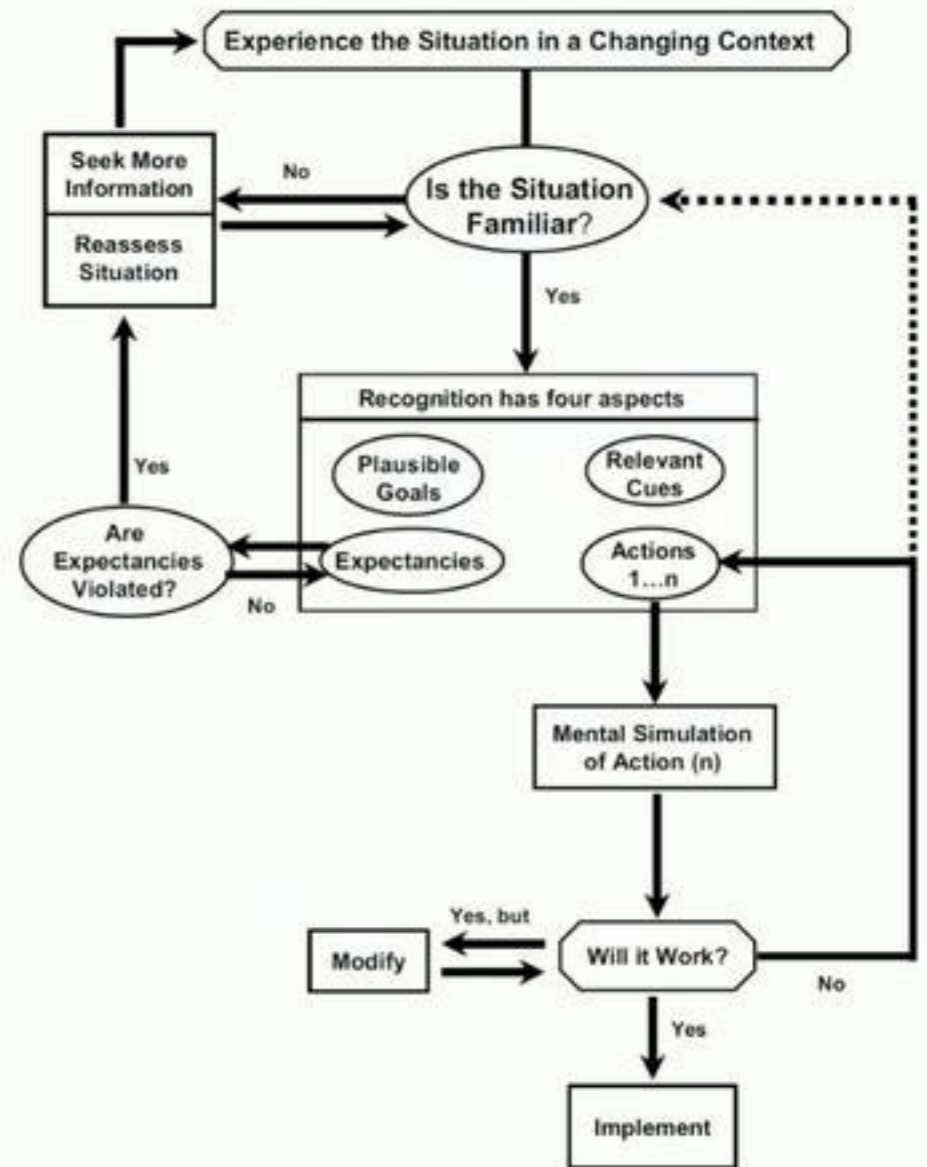
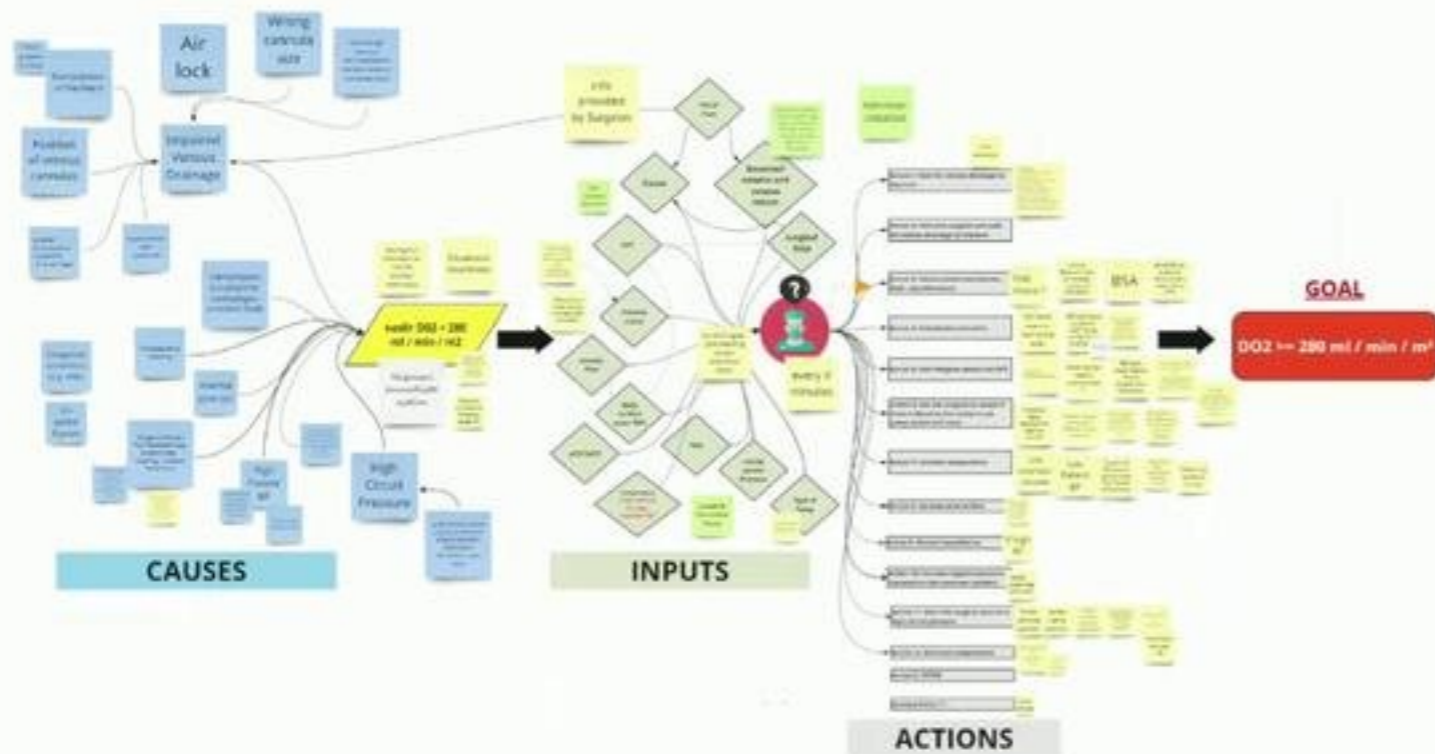
# Data Collection

- We interviewed with 8 perfusionists to understand how they make decision regarding managing level of DO<sub>2</sub>.
- Data collection in OR for 5 cases
- Working with 2 perfusionists to create xAI scenarios



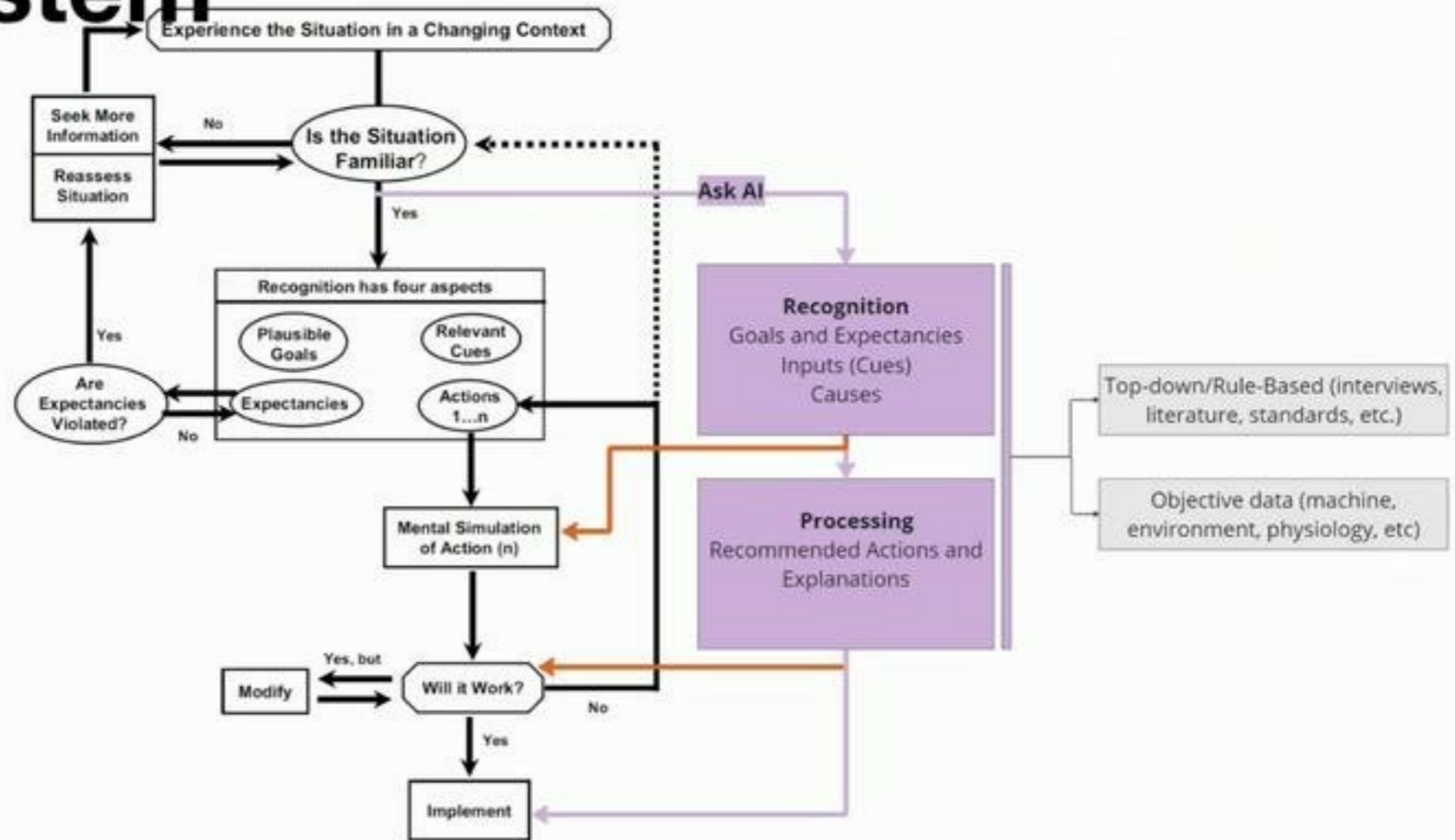






# Clinician-Centered xAI Decision Support System

- Instead of having AI to take over all decision making process, we are thinking to integrate different pieces of AI in part of human naturalistic decision process



# Next steps

- Fine tuning the framework and
- Using our perfusionists data collection (interview), to describe how the framework can be used.
- Conceptualize an envisioned decision support system in OR based on the framework
- Draft the paper

## xAI Techniques

- Decision Tree
- Counterfactual
- Case-based
- Feature Importance
- Nothing

# xAI Techniques

- Decision Tree
- Counterfactual
- Case-based
- Feature Importance
- Nothing

## Feature Importance

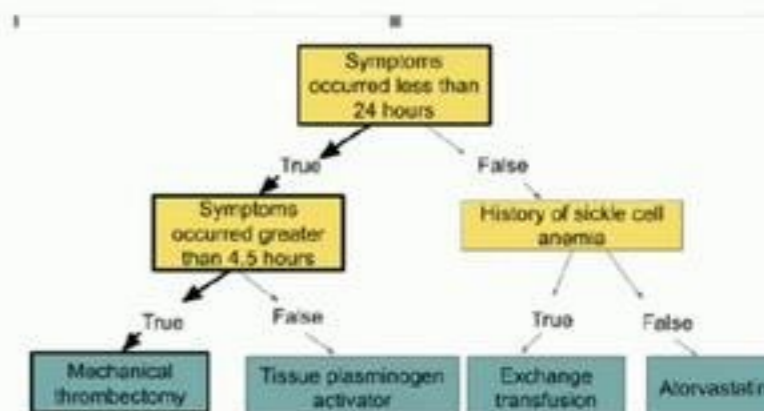
- A technique that involves identifying the most important features or factors that contributed to a particular decision or outcome.
- **Example:**

## Decision Tree

A 60 year old man with history of hypertension presents with a facial droop and right arm and leg weakness for six hours. He is also unable to speak and does not follow commands. His blood pressure is 155/80. He does not have a history of sickle cell anemia and he is not on an anticoagulant. Head CT shows no hemorrhage and CT angiography shows absence of contrast in the left middle cerebral artery. His low density lipoprotein (LDL) cholesterol is 190. Which of the following is the best next step?

Which of the following treatments is the best next step?

- Tissue plasminogen activator
- Mechanical thrombectomy
- Exchange transfusion
- Atorvastatin



Rate your agreement with the robot's suggestion

Strongly Disagree

Neutral

Strongly Agree

-100

0

100

I understand the  
reasoning behind the  
robot's suggestion

I agree with the  
robot's suggestion



**12** How confident are you in relation to your choice?

- ☐ Very confident
- ☐ Confident
- ☐ Moderately confident
- ☐ Very little confident



# 7 Aspects of xAI in Healthcare



Transparency



Domain Sense



Consistency



Generalizability



Generalizability



Trust/  
Performance



Fidelity

# Aspect 1: Transparency

## Transparency

Ability of the machine learning algorithm, model, and the features to be understandable by the user of the system.

## Admission Prediction

What is the likelihood of the patient being admitted to the hospital from the emergency department.



Katherine presents to the emergency department with severe headaches. She has multiple episodes of vomiting. She is evaluated by the clinical staff and has imaging and laboratory work done. She has very little medical history and considers herself active and healthy.

# Aspect 2: Domain Sense

## Domain Sense

The explanation should make sense in the domain of application and to the user of the system

## ED Census Prediction

Predict the number of patients in the emergency department (ED) at a given time

Several years later, Katherine revisits the emergency department due to abdominal pain. She has an elevated temperature and is dehydrated.

She is at the ED on a Friday after work. The ED is very crowded and she must wait several hours to be seen.

# Aspect 3: Consistency

## Consistency

The explanation should be consistent across different models and across different runs of the model

## LWBS

Left without being seen refers to a patient leaving the facility without being seen by a physician

~18:00

Patients at Risk of Leaving Without Being Seen

Family Name	Given Name	Age	Chief Complaint	PCP Name	Insurance Type	Checked In Time	Elapsed Waiting Time	Past LWBS Occurences	Predicted LWBS Score	Factors
Franklin	Samuel	56	WOUND INFECTION	Overman	Private	18:30	90	2	62	Prior LWBS
Pierce	Jerome	41	ABDOMINAL PAIN	Mark	Uninsured	17:45	135	1	40	Insurance status
Collins	Katherine	37	ABDOMINAL PAIN	Monroe	Private	16:52	68	0	24	Past history

~22:00

Patients at Risk of Leaving Without Being Seen

Family Name	Given Name	Age	Chief Complaint	PCP Name	Insurance Type	Checked In Time	Elapsed Waiting Time	Past LWBS Occurences	Predicted LWBS Score	Factors
Franklin	Samuel	56	WOUND INFECTION	Overman	Private	18:30	90	2	62	Temperature
Pierce	Jerome	41	ABDOMINAL PAIN	Mark	Uninsured	17:45	135	1	40	Insurance status
Collins	Katherine	37	ABDOMINAL PAIN	Monroe	Private	16:52	314	0	24	Chief complaint

# Aspect 4: Parsimony

## Parsimony

The explanation should be as simple as possible

Applies to both the complexity of the explanation and the number of features provided to explain

## Admission Disposition

Where in the hospital the patient should go once they are admitted



# Aspect 4: Parsimony

- MDL (Minimum Description Length) and Occam's Razor
- **Occam's Razor:** To derive a unifying diagnosis that can explain all of a patient's symptoms
- **Hickam's Dictum:** A man can have as many diseases as he damn well pleases
- Occam's Razor in Machine Learning [Domingos 1999]
  - Occam's First Razor
  - Occam's Second Razor
- The simplest explanation is not always the best explanation



© KenSci., Inc. 2018.

# Aspect 5: Generalizability

## Generalizability

Models and explanations should be generalizable across problem whenever possible

Katherine eventually develops diabetes. It is well controlled and she takes her medications as directed. One afternoon, she is admitted from clinic due to highly elevated glucose levels and a urinary tract infection. Her nurse tells her that based on her illness and other factors, her predicted length of stay is 3 days.

## Length of Stay

The time that a patient will spend at a particular healthcare facility

Room Number	Patient ID	Last Name	First Name	Age	Gender	Attending Provider	Chief Complaint	Admission Date	Predicted Discharge Date	Elapsed LOS	Predicted LOS	Explanation
9B	3408738	Mascroft	Pete	53	Male	Thomas Louwers	Dyspnea	5/22/2018	5/28/2018	6	6	Diabetes
15A	20150155	Hookano	Karina	24	Female	Jonathan Miller	Anemia	5/28/2018	5/31/2018	0	3	"
7C	26058755	Grant	Tien	71	Male	Thomas Louwers	SOB	5/24/2018	5/28/2018	4	4	Diabetes
10A	4664428	Bercier	Carmelia	32	Female	Venessa Overman	Fatigue	5/26/2018	5/29/2018	2	3	"
14A	1985950	Brea	Tameika	46	Female	Susan Mark	Sepsis	5/27/2018	5/30/2018	1	3	"
3B	8354235	Lico	Lien	53	Male	Susan Mark	Abd pain	5/22/2018	6/1/2018	6	10	Diabetes
13A	4205535	Sayergh	Vina	55	Female	Jonathan Miller	Fatigue	5/26/2018	5/31/2018	2	5	"
11C	42399976	Zeidan	Oma	68	Female	Susan Mark	OKD	5/28/2018	5/30/2018	0	2	Diabetes
12C	59204677	Rile	Joanne	61	Female	Venessa Overman	UTI	5/25/2018	5/29/2018	3	4	"

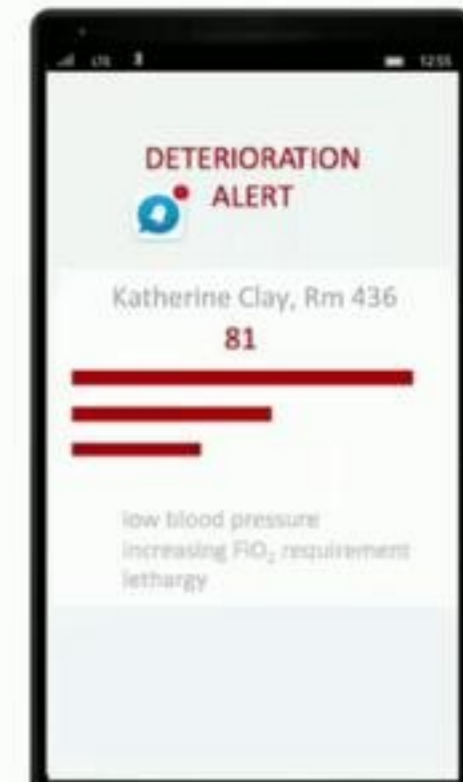
# Aspect 6: Trust/Performance

## Trust / Performance

- The expectation that the corresponding predictive algorithm for explanations should have a certain performance

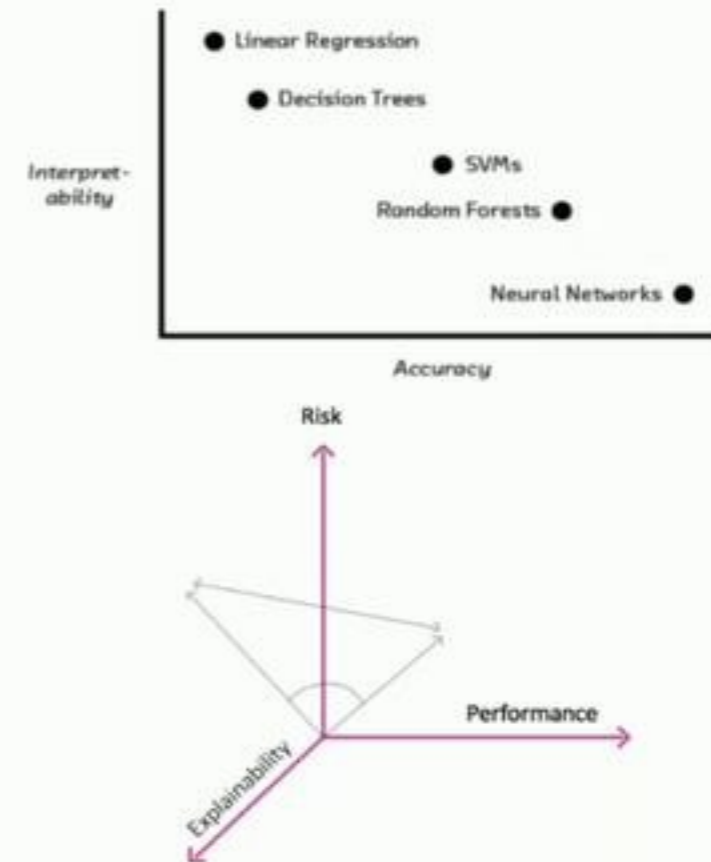
## ICU Transfer Prediction

- Predict if a patient on the hospital ward will require transfer to the intensive care unit due to increasing acuity of care needs



# Aspect 6: Trust/Performance

- Expectation that the predictive system has a sufficiently high performance e.g., precision, recall, AUC etc. [Lipton 2016, Hill 2018]
- Explanations accompanied with sub-par predictions can foster distrust
- The model should perform sufficiently well on the prediction task in its intended use
- The model has at least parity with the performance of human practitioners
- Trauma patients: vital signs and lab criteria fulfill criteria to trigger alarm, leads to increasing numbers of false positives [Nguyen 2014]



© KenSci., Inc. 2018.

# Aspect 7: Fidelity

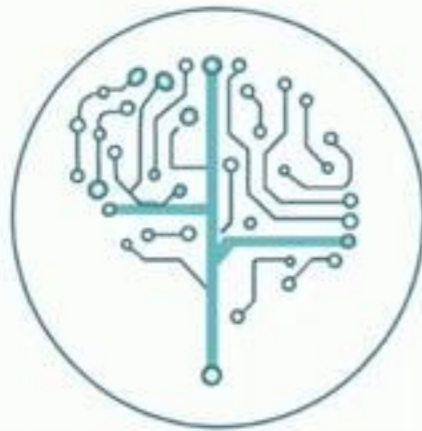
## Fidelity

- The expectation that the explanation and the predictive model align well with one another

## Risk of Readmission

- Predict if the patient will be readmitted within a particular span in time, i.e. 30 days from time of discharge





### Artificial Intelligence

Custom machine learning model trained on clinical data



### Explanations

Provides interpretable insights and justifications



### Clinical Expertise

Input and oversight of clinicians and healthcare professionals

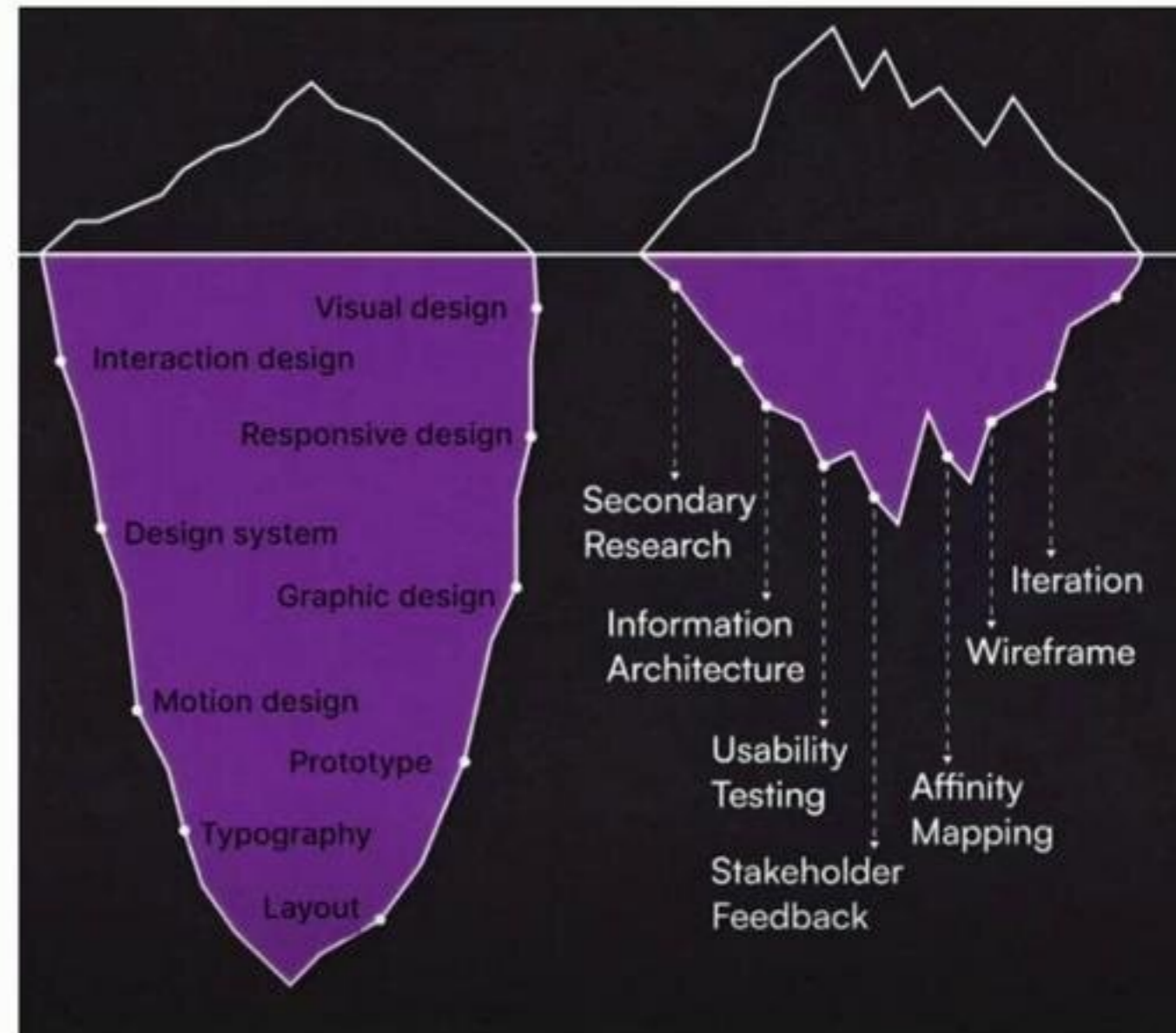


### Assistive Intelligence

Enhances and augments clinical decision-making

# Human-Centered XAI Principles

- Treat explanation as UX, not afterthought.
- Adapt depth to user expertise (nurse vs data scientist).
- Support dialogue: ask-answer loops.

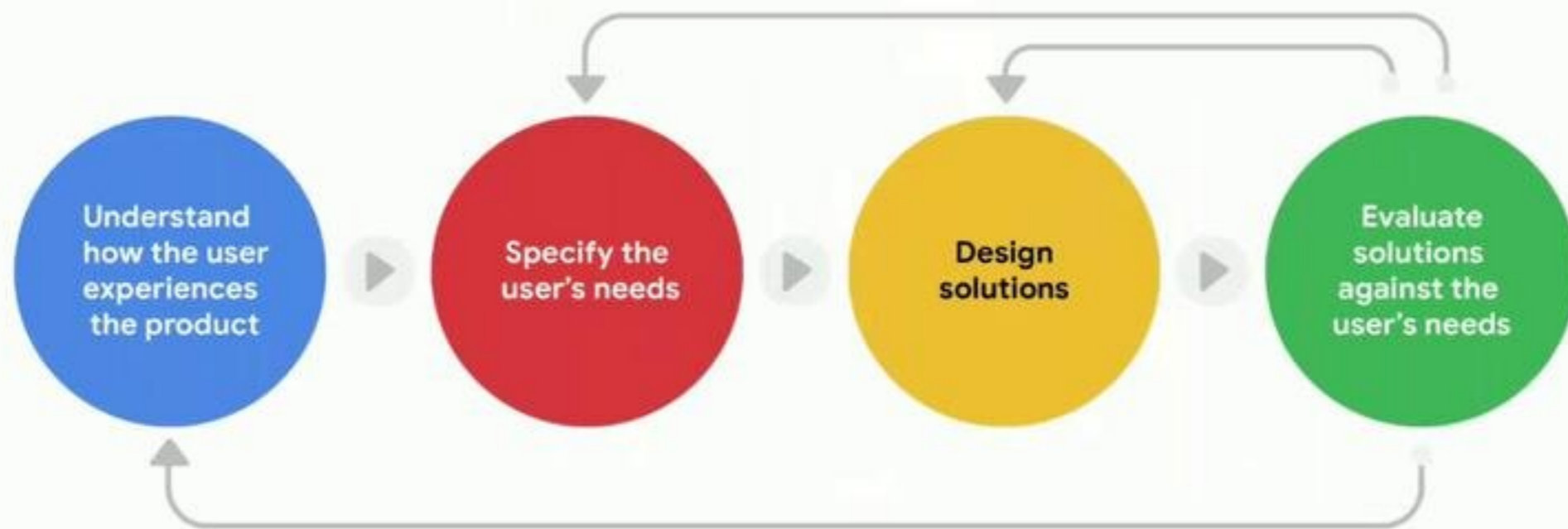


# Design Challenge: No One-Size-Fits-All



# Question-Driven Design Framework

- Collect user questions (Why? How to improve?).
- Map to explanation types & visuals.
- Iterate with feedback sessions.



# Step 1: Gather Clinician Questions

- Card-sorting workshops & shadowing.
- Log real-time questions during rounds.
- Prioritize by frequency & decision impact.



## Step 2: Analyze & Prioritize



Cluster into categories: Why, What-if, Performance.



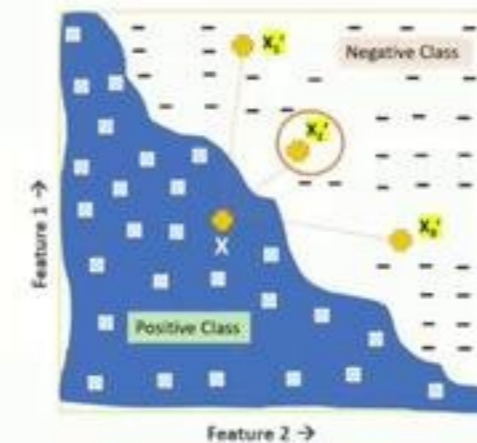
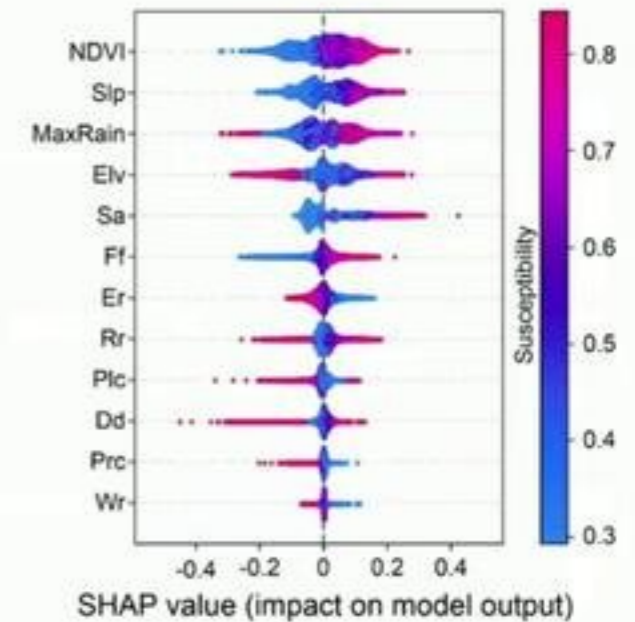
Identify unmet needs & misconceptions.



Translate into user requirements.

## Step 3: Map to Techniques

- Why → SHAP, rule lists.
- What-if → PDP, counterfactual.
- Performance → calibration, confidence intervals.



Chrome File Edit View History Bookmarks Profiles Tab Window Help

GoTo Meeting

app.goto.com/meeting/333782365

PATRICIA E SORTILLON G Everyone

Conference Support

Rayan Harari

PG

You're sharing your screen

PG

Conference Support

PATRICIA E SORTILLON G

Record React Mic Camera Share Leave Captions Pop out

26

zoom

Chrome File Edit View History Bookmarks Profiles Tab Window Help

GoTo Meeting

app.goto.com/meeting/333782365

New Chrome available

All Bookmarks

Rayan Harari, PATRICIA E SO...

Everyone

Conference Support

Rayan Harari

PG

PG

PATRICIA E SORTILLON G

Record React Mic Camera Share Leave Captions Pop out

26

Zoom

WhatsApp

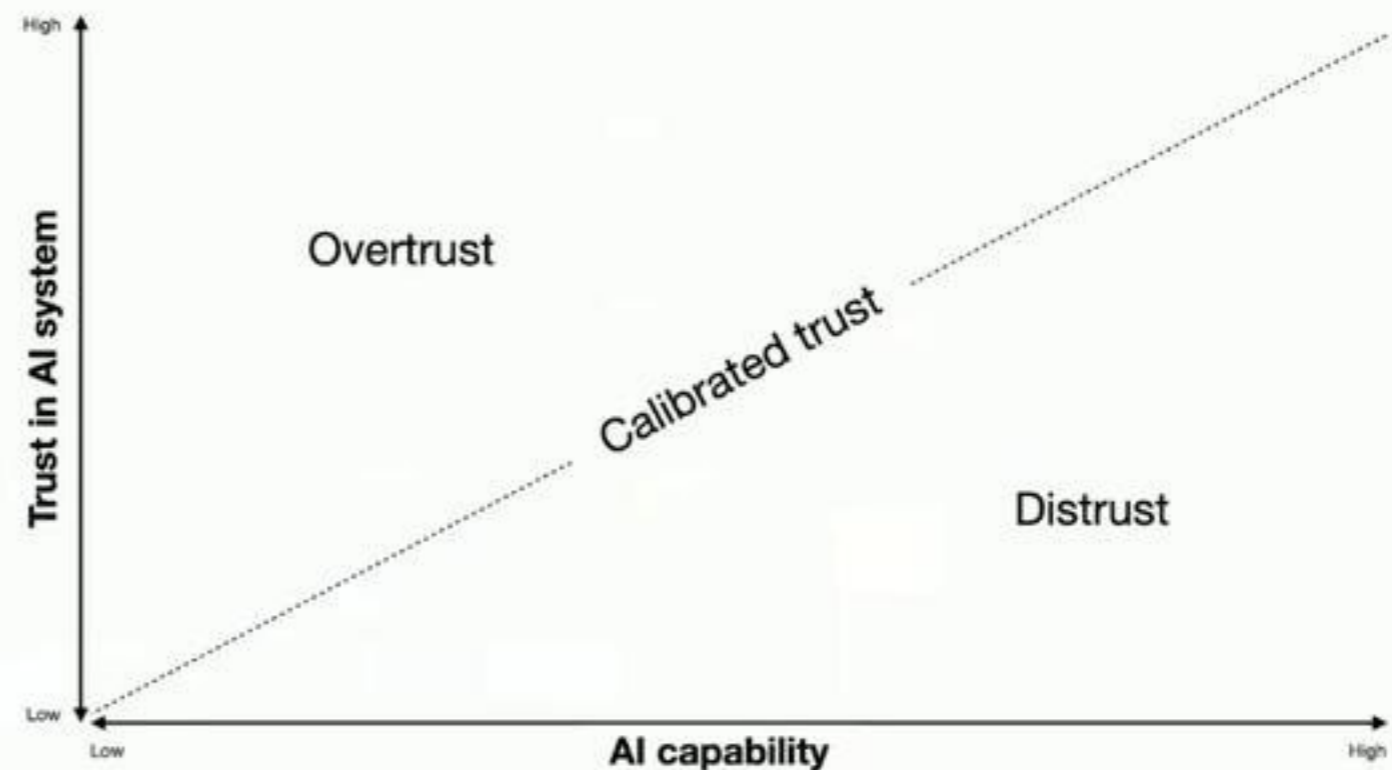
# Bridging Algorithm-UX Gap

- Data scientists embed tags (units, ranges) for UI.
- Designers ensure colorblind-safe palettes.
- Joint design reviews avoid misinterpretation.



# Mitigating Over-Trust & Cognitive Load

- Provide uncertainty & encourage double-check.
- Use cognitive forcing functions (Bućinca 2021).
- Limit number of explanatory elements per alert.



# Key Clinician Questions

- 'Why high risk?'; 'How to reduce it?'; 'Is model confident?'.
- Dashboard groups answers accordingly.
- Fast access crucial during 3-minute bedside encounters.

# The 4 Levels of Evaluation

Level 1: Technical  
performance

Level 2: Task-level utility

Level 3: Clinical workflow  
integration

Level 4:  
Sociotechnical/contextual  
impact

# Explainability $\neq$ Interpretability

Explanation =  
how it behaves  
in context

Interpretability  
= how it's built

Need both for  
safety and  
trust

# What Makes an Explanation Useful?

- Relevant to current decision context
- Reflects model reasoning clearly
- Supports, not replaces, human judgment

# From Publication to Practice

- Successful tools: sepsis early warning, imaging triage
- Characteristics: iterative dev, strong user feedback
- Barriers: lack of generalizability, alert fatigue

# Summary: Key Lessons Learned

- AUC, etc is not enough—context matters
- Evaluation must go beyond model to system
- Design must center users, workflows, and equity

# Next Steps for Learners

- Use the framework to critique published models
- Involve users early in your AI projects
- Design evaluations that reflect real-world use

# Human-Centered and Explainable AI: Foundations and Applications in Medicine

**Rayan Ebnali Harari, PhD**

Harvard Medical School | MGB