NANYANG
TECHNOLOGICAL
UNIVERSITY
SINGAPORE

Explainable AI

Yu Han

han.yu@ntu.edu.sg

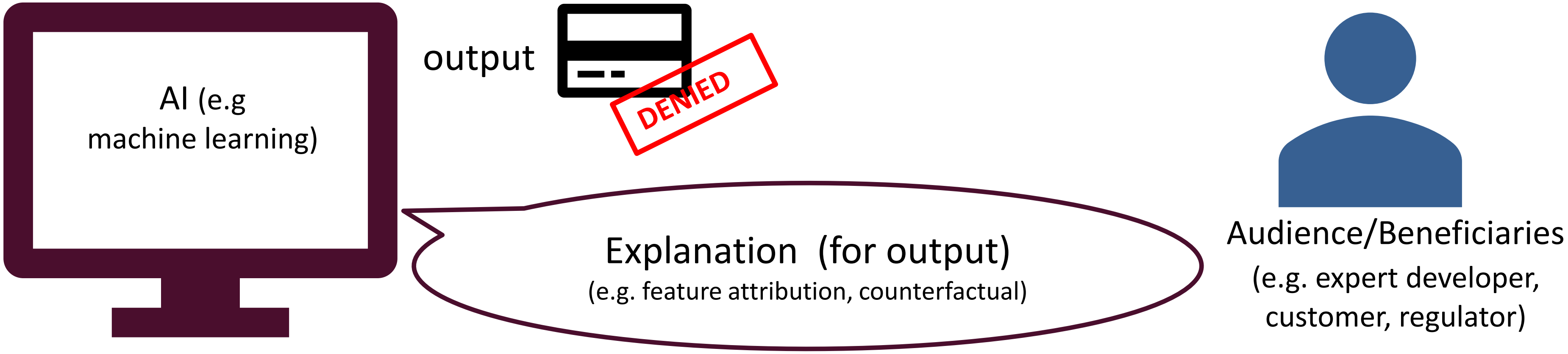
Nanyang Assistant Professor
School of Computer Science and Engineering
Nanyang Technological University

GDPR again …

> “
>
> *Companies should commit to ensuring systems that could fall under GDPR, including AI, will be compliant. The threat of* ***sizeable fines of €20 million or 4% of global turnover*** *provides a sharp incentive.*
>
> *Article 22 of GDPR empowers individuals with the* ***right to demand an explanation of how an AI system made a decision that affects them.***
>
> ”

- European Commision

XAI



AI (e.g
machine learning)

output

Explanation (for output)
(e.g. feature attribution, counterfactual)

Audience/Beneficiaries
(e.g. expert developer,
customer, regulator)

Explanation as feature attribution:

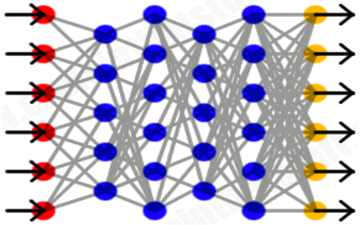
Card denied because client is **credit unworthy**, despite **good salary**

Counterfactual explanation:

Had the client had **a good credit score** the card would have been granted

XAI

AI System



http://explainthatstuff.com

http://listverse.com/
©2007–2017 Listverse Ltd

Transportation

https://www.re-work.co/
© 2017 RE-WORK X LTD

Finance

http://listverse.com/
©2007–2017 Listverse Ltd

Security

http://www.bbc.com/
© 2017 BBC
Credit: Getty Images

Legal

http://blog.soliant.com/
©2004-2017 Soliant Health

Medicine

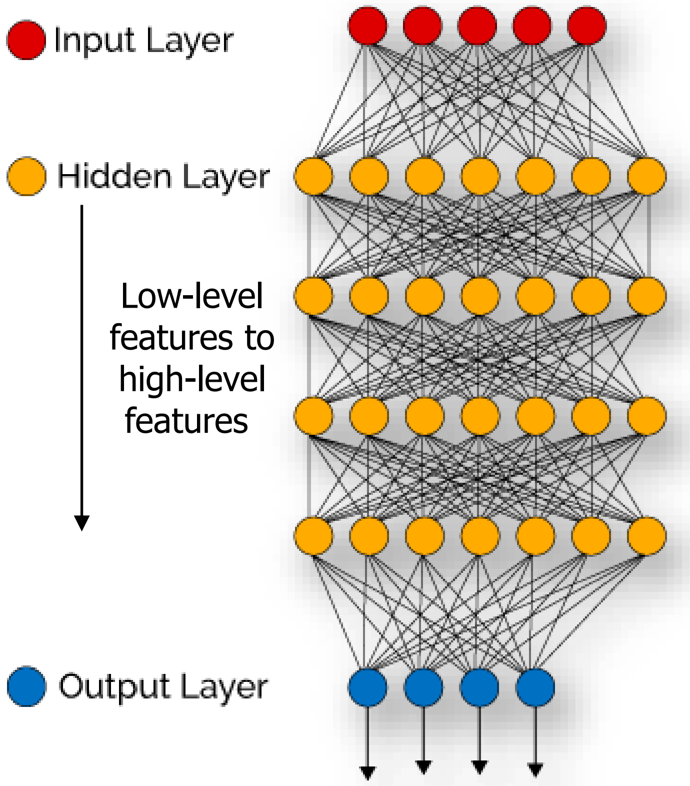
https://futureoflife.org/
©FLI - Future of Life Institute

Military

User

- 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?

Deep Learning Working Principles

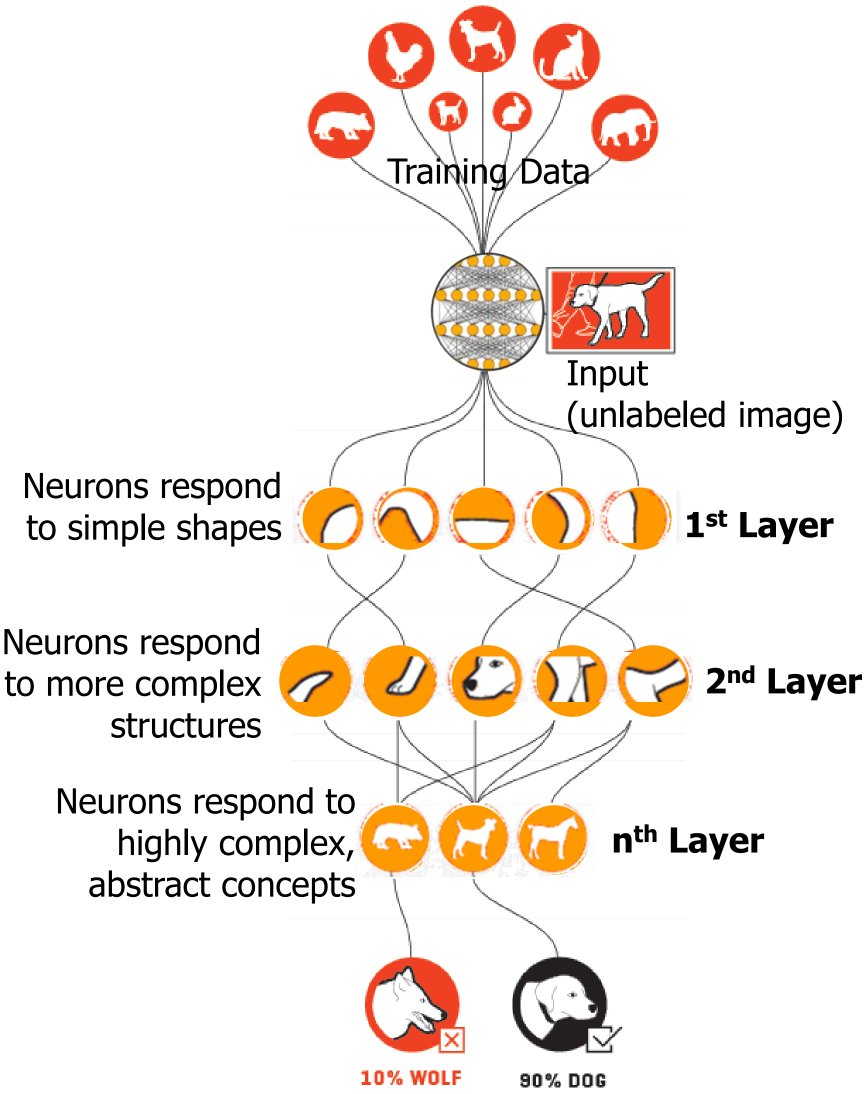


● Input Layer

● Hidden Layer

Low-level features to high-level features

● Output Layer



Training Data

Input (unlabeled image)

Neurons respond to simple shapes

1st Layer

Neurons respond to more complex structures

2nd Layer

Neurons respond to highly complex, abstract concepts

nth Layer

10% WOLF

90% DOG

XAI

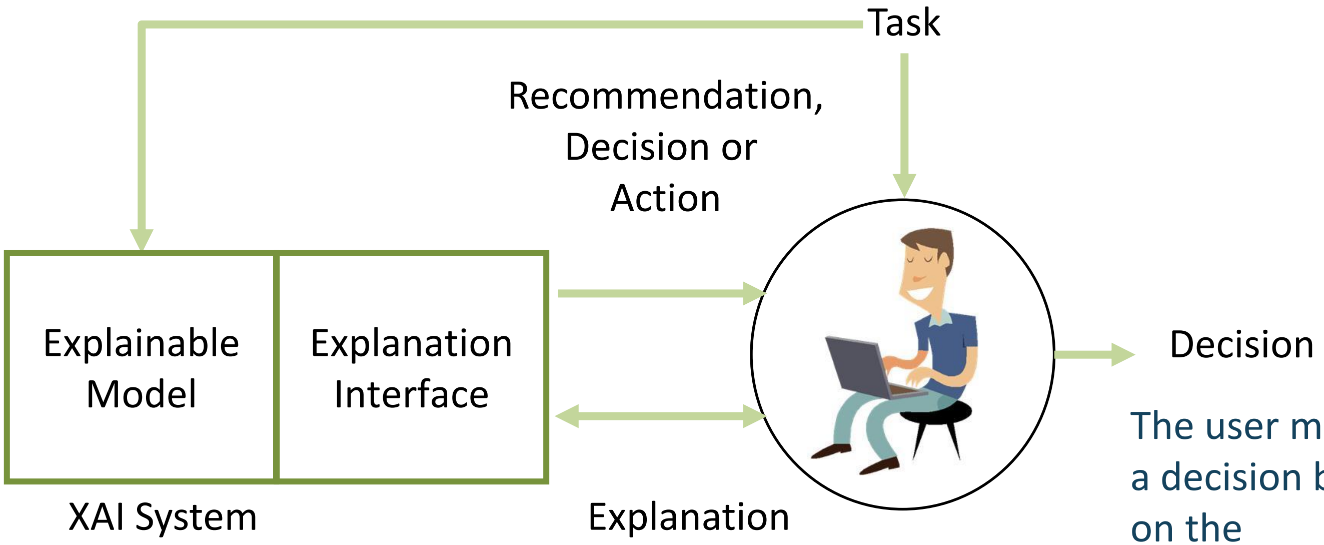
“The function of reasoning is … to devise and evaluate arguments intended to persuade.”

- Mercier, Sperber: BEHAVIORAL AND BRAIN SCIENCES (2011)

“looking at how humans explain to each other can serve as a useful starting point for explanation in artificial intelligence”

- Tim Miller AIJ2019

XAI Framework



Task

Recommendation, Decision or Action

Explainable Model

Explanation Interface

XAI System

Explanation

Decision

The system takes input from the current task and makes a recommendation, decision, or action

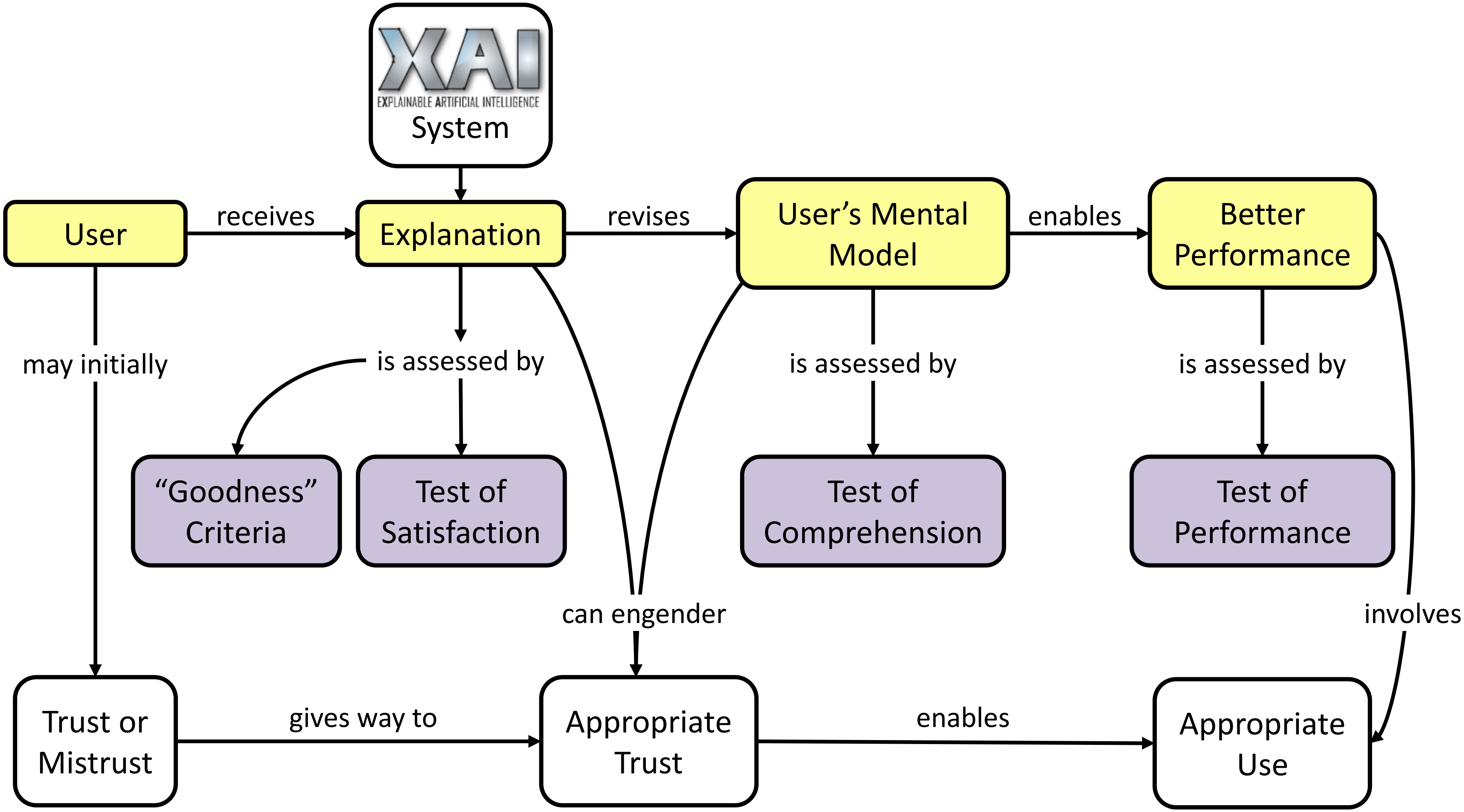
The system provides an explanation to the user that justifies its recommendation, decision, or action

The user makes a decision based on the explanation

XAI Framework

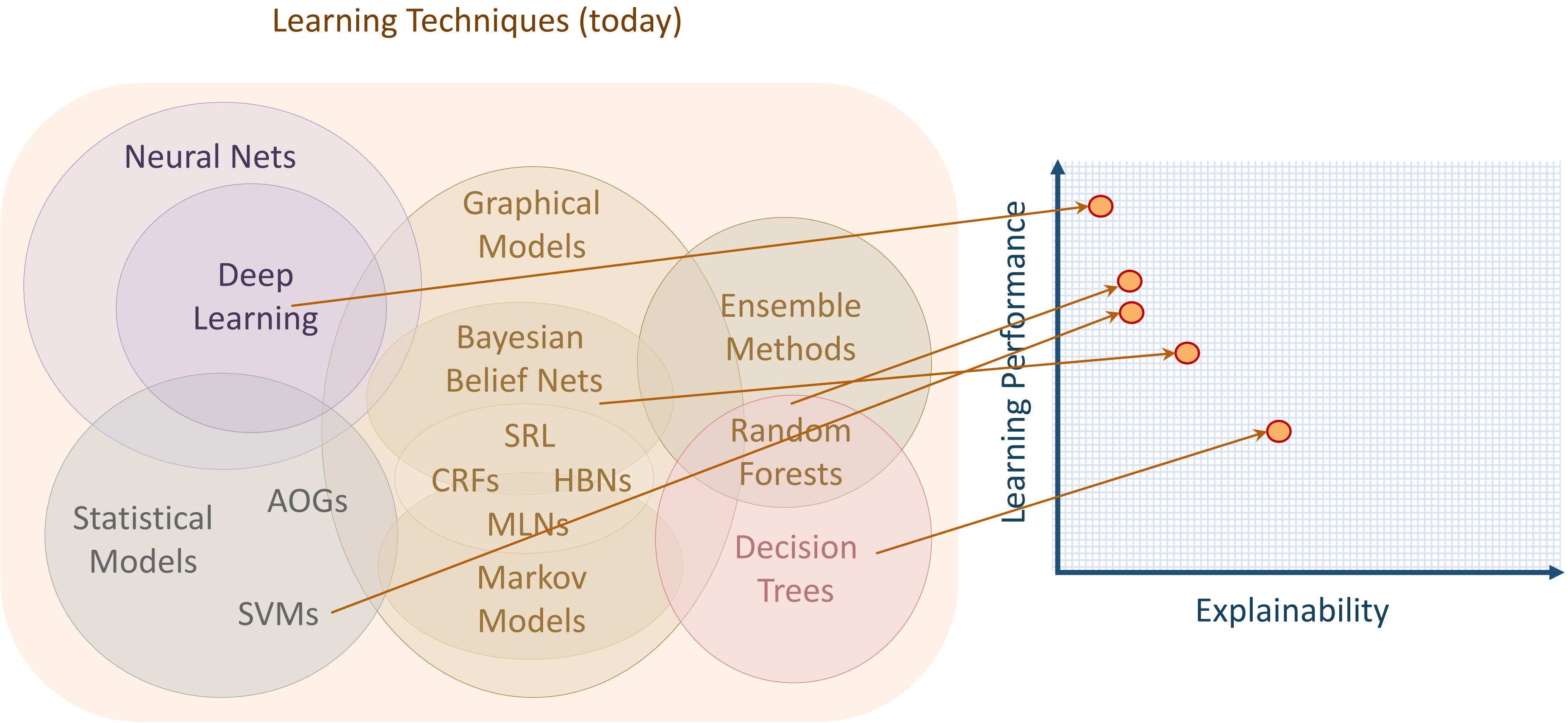
XAI Process

XAI Metrics



http://www.darpa.mil/

Performance vs. Explainability



Terminology

Understandability

- **Understandability** (or **intelligibility**) refers to the characteristic of a model to make a human understand its function – how the model works – without any need for explaining its internal structure or the algorithmic means by which the model processes data internally

Comprehensibility

- **Comprehensibility:** when conceived for machine learning models, comprehensibility refers to the ability of a learning algorithm to represent its learned knowledge in a human understandable fashion

Interpretability

- **Interpretability:** it is defined as the ability to explain or to provide the meaning in understandable terms to a human.

Explainability

- **Explainability:** it is associated with the notion of explanation as an interface between humans and a decision maker
  - that is, at the same time, both an accurate proxy of the decision maker and comprehensible to humans

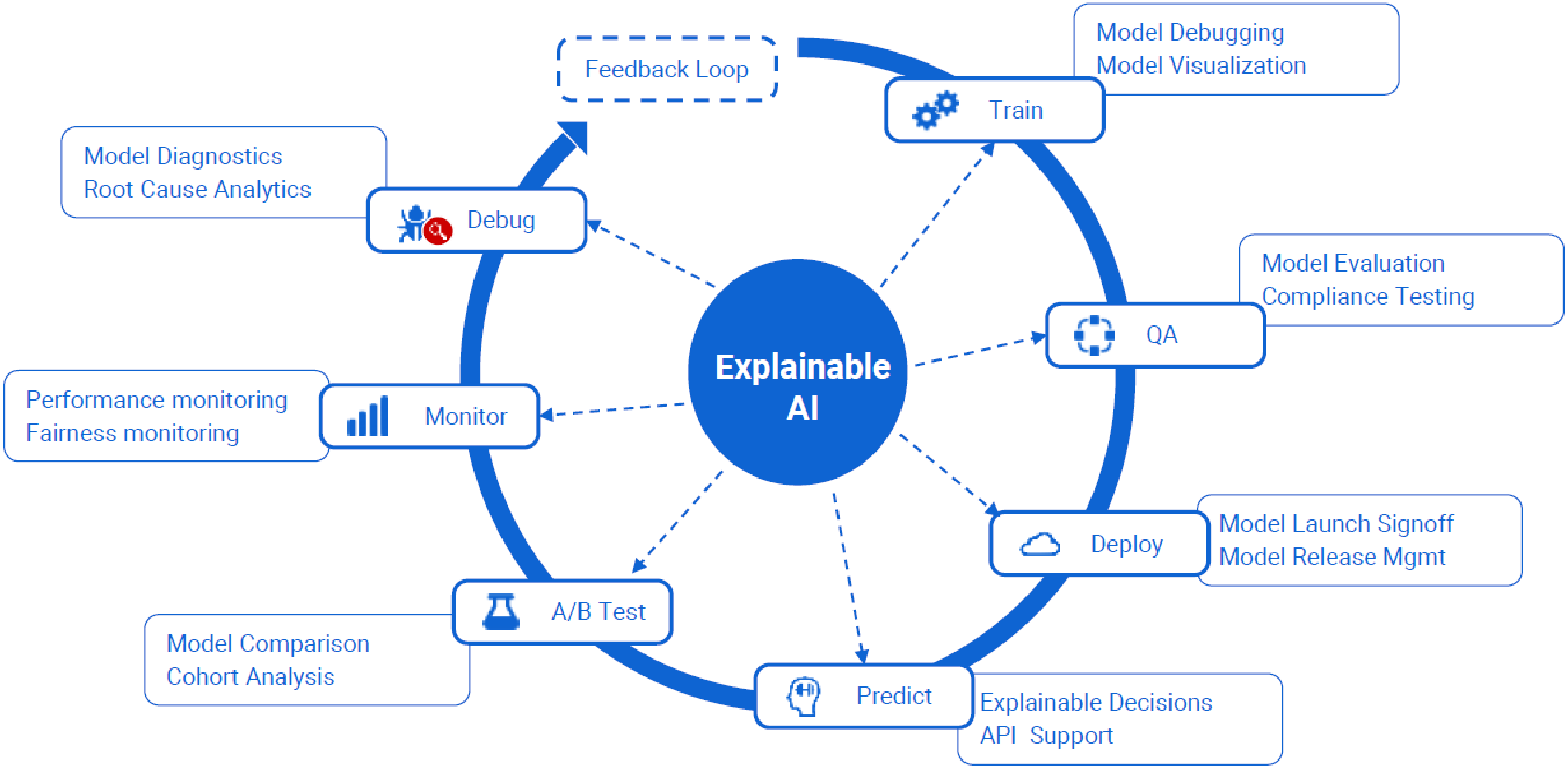
Explicability

- **Explicability:**
  - Making AI decisions obvious to a human being (i.e. a human being can understand the reason behind an AI decision without explanation)
  - Might not be the optimal solution!

Transparency

- **Transparency:** a model is considered to be transparent if by itself it is understandable. A model can feature different degrees of understandability.

The Use of XAI

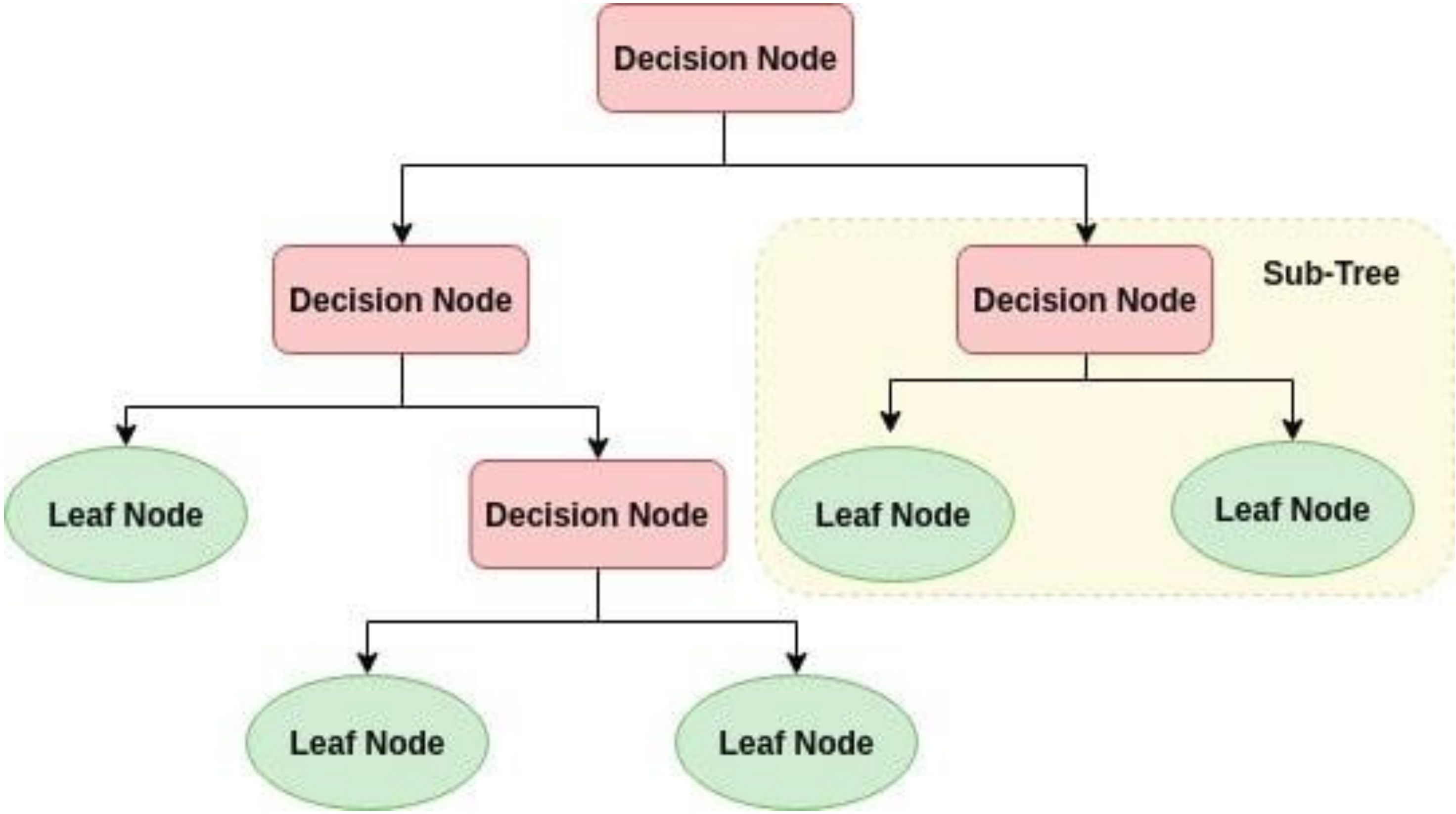


Measuring Explanation Effectiveness

Measure of Explanation Effectiveness
User Satisfaction
• Clarity of the explanation (user rating) • Utility of the explanation (user rating)
Mental Model
• Understanding individual decisions • Understanding the overall model • Strength/weakness assessment • 'What will it do' prediction • 'How do I intervene' prediction
Task Performance
• Does the explanation improve the user's decision, task performance? • Artificial decision tasks introduced to diagnose the user's understanding
Trust Assessment
• Appropriate future use and trust
Correctability
• Identifying errors • Correcting errors • Continuous training

Building an Explainable Model (Decision Tree)

The Basics of Decision Tree



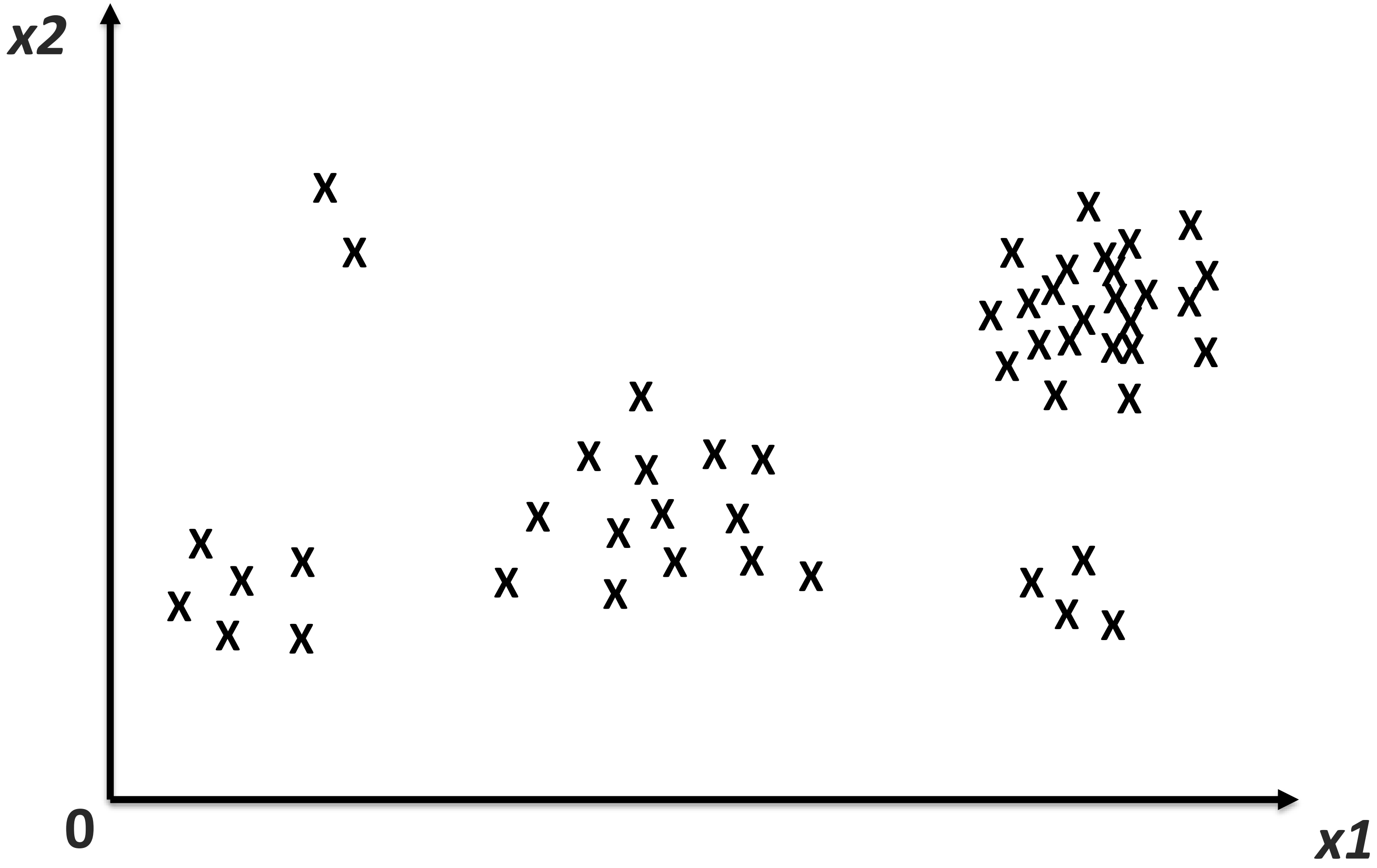
The Basics of Decision Tree

- A Decision Tree is a tree-structured plan of a set of attributes to test in order to predict the output.
- A type of supervised learning approaches
- Mostly used in classification problems
- Good interpretability / visualizability
- Not the best performance

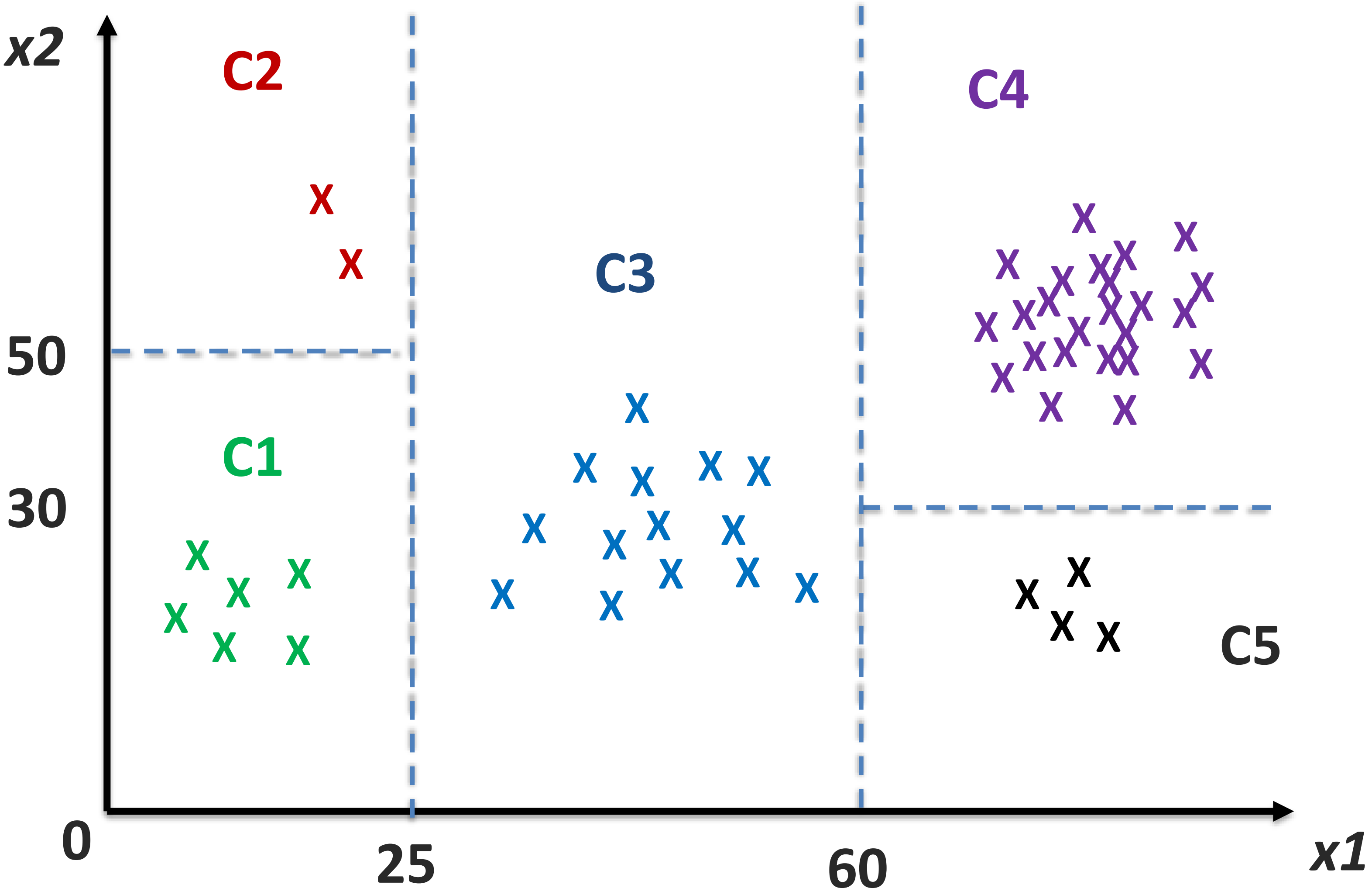
Terminology

1. **Root Node (Top Decision Node):** It represents the entire population and can be further divided into two or more homogeneous sets.
2. **Splitting:** It is a process of dividing a node into two or more sub-nodes.
3. **Decision Node:** When a sub-node splits into further sub-nodes, then it is called a decision node.
4. **Leaf/ Terminal Node:** Nodes with no children (no further split) is called Leaf or Terminal node.
5. **Pruning:** When we reduce the size of decision trees by removing nodes (opposite of Splitting), the process is called pruning.
6. **Branch / Sub-Tree:** A sub section of the decision tree is called branch or sub-tree.
7. **Parent and Child Node:** A node, which is divided into sub-nodes is called a parent node of sub-nodes whereas sub-nodes are the child of a parent node.

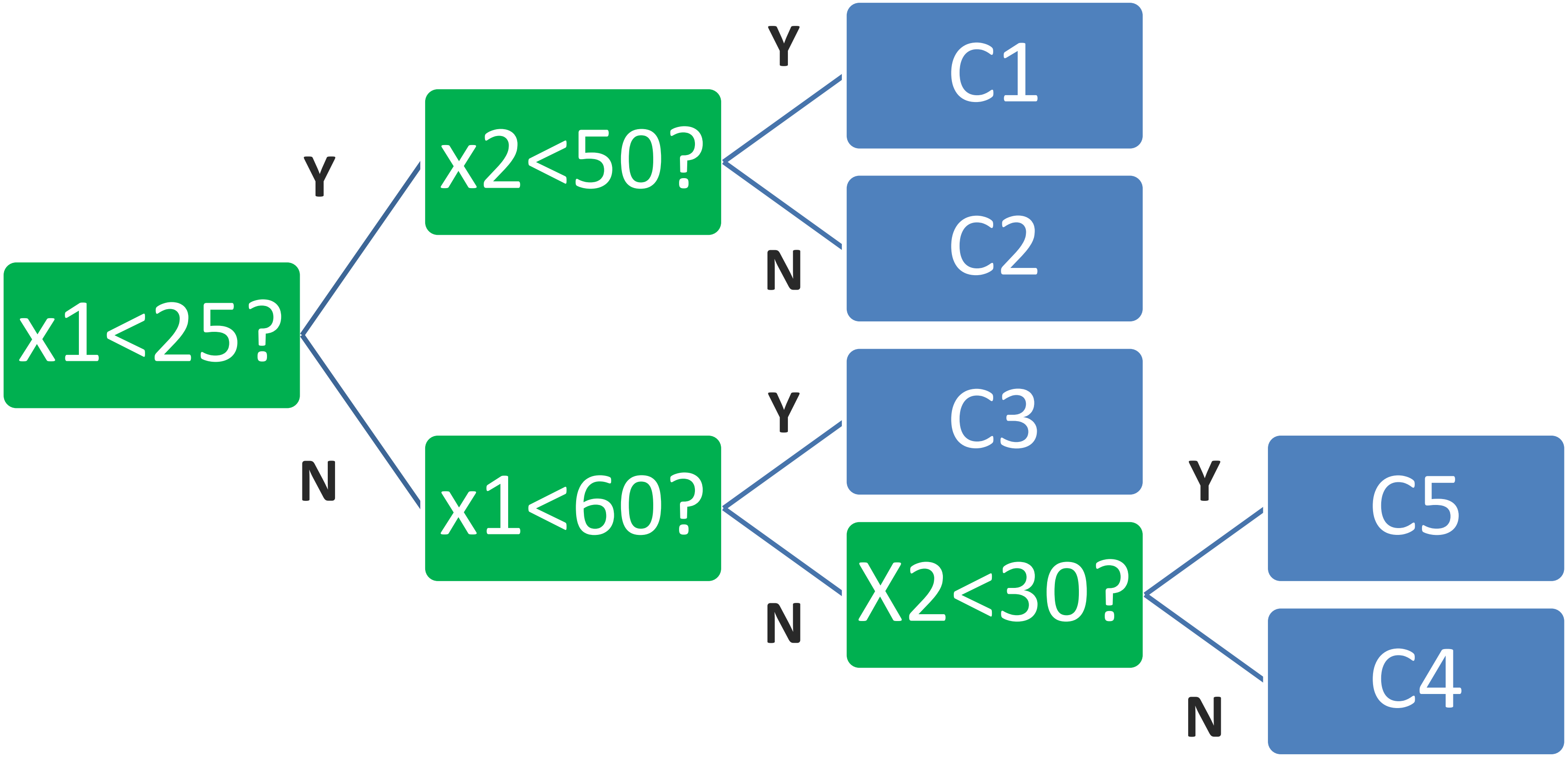
Decision Tree Example



Decision Tree Example



Decision Tree Example



Discussions

- Pro:
  - Very good interpretability
- Con:
  - If the boundaries between classes are not crisp (most real-world applications will fall into this category), the predictions by decisions trees can be inaccurate.

Explainable AI

Yu Han

han.yu@ntu.edu.sg

Nanyang Assistant Professor
School of Computer Science and Engineering
Nanyang Technological University