CSCI 4800/5800
Explainable AI

Explainable AI (XAI):
Explainability and Interpretability

Denver

Definitions of "explanation" and "interpretation" found in XAI literature

Source	Explanation	Interpretation
(Lewis, 1986)	"someone who is in possession of some information about the causal history of some event (. . .) tries to convey it to someone else."	-
(Josephson & Josephson, 1996)	"assignment of causal responsibility"	-
(Lombrozo, 2006)	"central to our sense of understanding and the currency in which we exchange beliefs. Explanations often support the broader function of guiding reasoning."	-
(Biran & Cotton, 2017)	-	"the degree to which an observer can understand the cause of a decision"
(Lundberg & Lee, 2017)	"interpretable approximation of the original [complex] model"	-
(Montavon et al., 2018)	"collection of features of the interpretable domain, that have contributed for a given example to produce a decision (e.g., classification or regression)"	"mapping of an abstract concept (e.g., a predicted class) into a domain that the human can make sense of"

Definitions of "explanation" and "interpretation" found in XAI literature (continued)

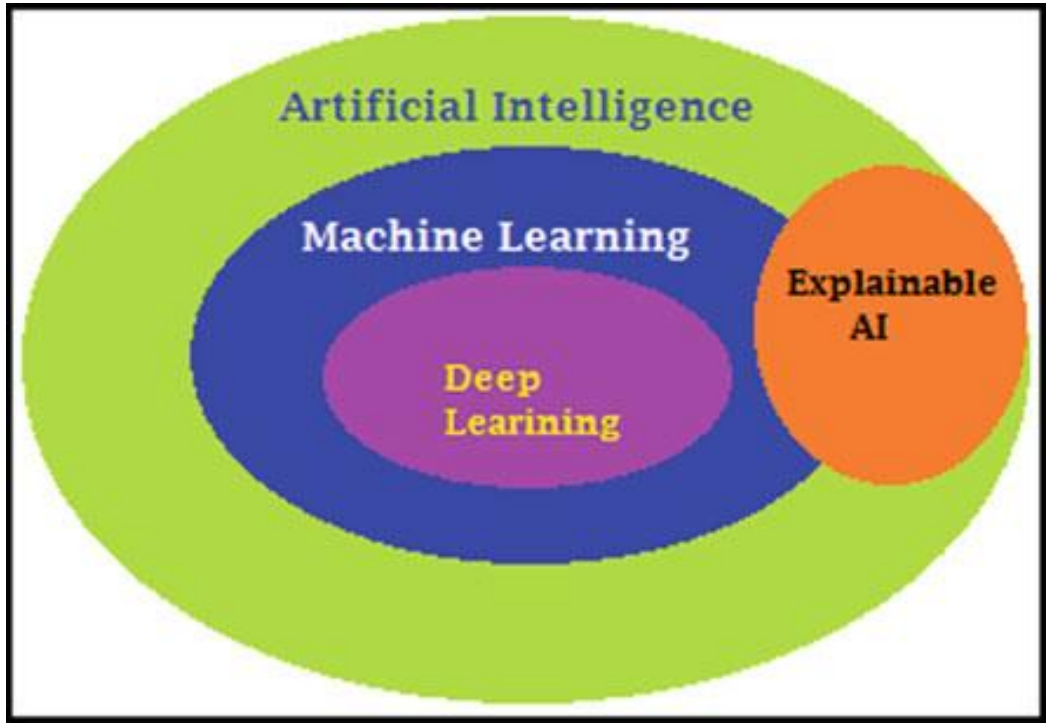
Source	Explanation	Interpretation
(Dam et al., 2018)	"measures the degree to which a human observer can understand the reasons behind a decision (e.g., a prediction) made by the model"	-
(Doshi-Velez & Kim, 2018)	-	"to explain or to present in understandable terms to a human"
(Lakkaraju et al., 2019)	-	"quantifies how easy it is to understand and reason about the explanation. Depends on the complexity of the explanation"
(Vilone & Longo, 2020)	"the collection of features of an interpretable domain that contributed to produce a prediction for a given item"	"the capacity to provide or bring out the meaning of an abstract concept"
(Schmid & Finzel, 2020)	"in human–human interaction, explanations have the function to make something clear by giving a detailed description, a reason, or justification"	-
(Al-Shedivat et al., 2020)	"local approximation of a complex model [by another model]"	-

Definitions of "explanation" and "interpretation" found in various dictionaries

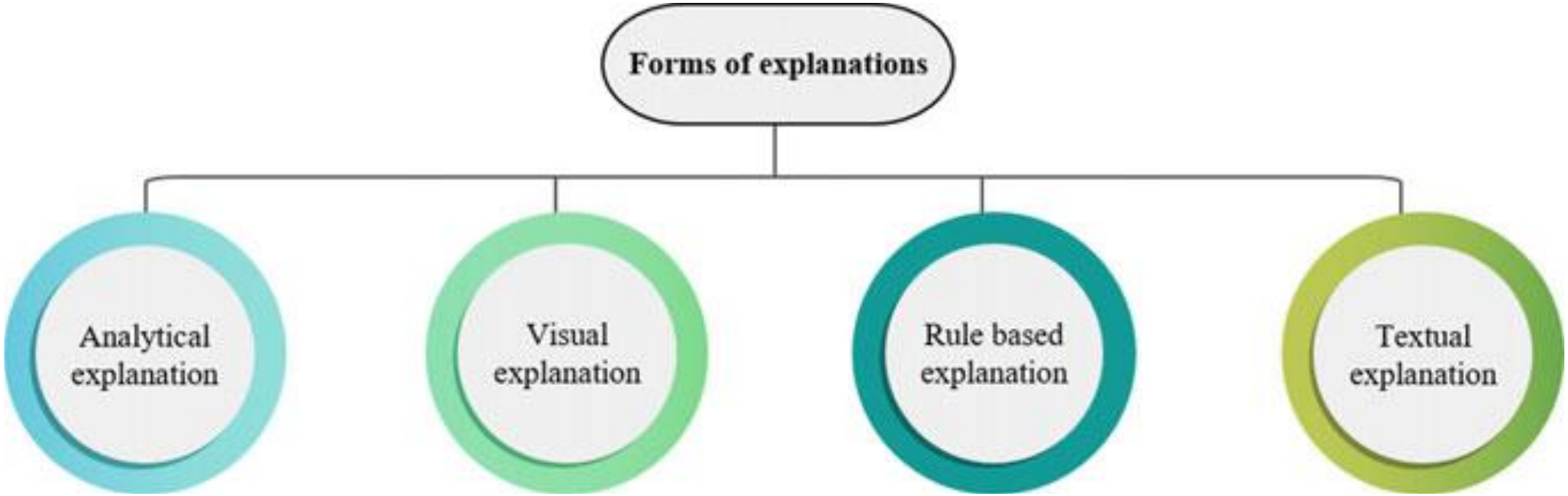
Source	Explanation	Interpretation
Merriam-Webster	**act** of making plain or understandable	**action to explain** or tell the meaning of
Cambridge	**the details** or other information that someone gives to make something clear or easy to understand	**an explanation** or **opinion** of what something means
Oxford	**a statement** or **account** that makes something clear	**the action of explaining** the meaning of something
Dictionary.com	**statement** made to clarify something and make it understandable	**explain**; action to give or provide the meaning of; explicate; elucidate
Princeton	**statement** that makes something comprehensible by describing the relevant structure or operation or circumstances etc.	**an explanation** of something that is not immediately obvious; a mental representation of the meaning or significance of something
Wikipedia	**a set of statements** usually constructed to describe a set of facts that clarifies the causes, context, and consequences of those facts	**A philosophical interpretation** is the assignment of meanings to various concepts, symbols, or objects under consideration.

Relationships among AI, ML. DL, and XAI

- Artificial Intelligence (AI)
- Machine Learning (ML)
- Deep Learning (DL)
- Explainable AI (XAI)



Different forms of explanation useful for XAI



Example of analytical explanation

- Analytical explanations are generated by measuring the contribution of the input features to the model's outcome.
- They are represented by various numeric metrics such as saliency, causal importance, feature importance, features confidence score, and mutual importance.
- Domain experts mostly utilize them to view and explore the data concerning their feature importance.

Text record: Where is mile high stadium?

Prediction: ***LOC: other***

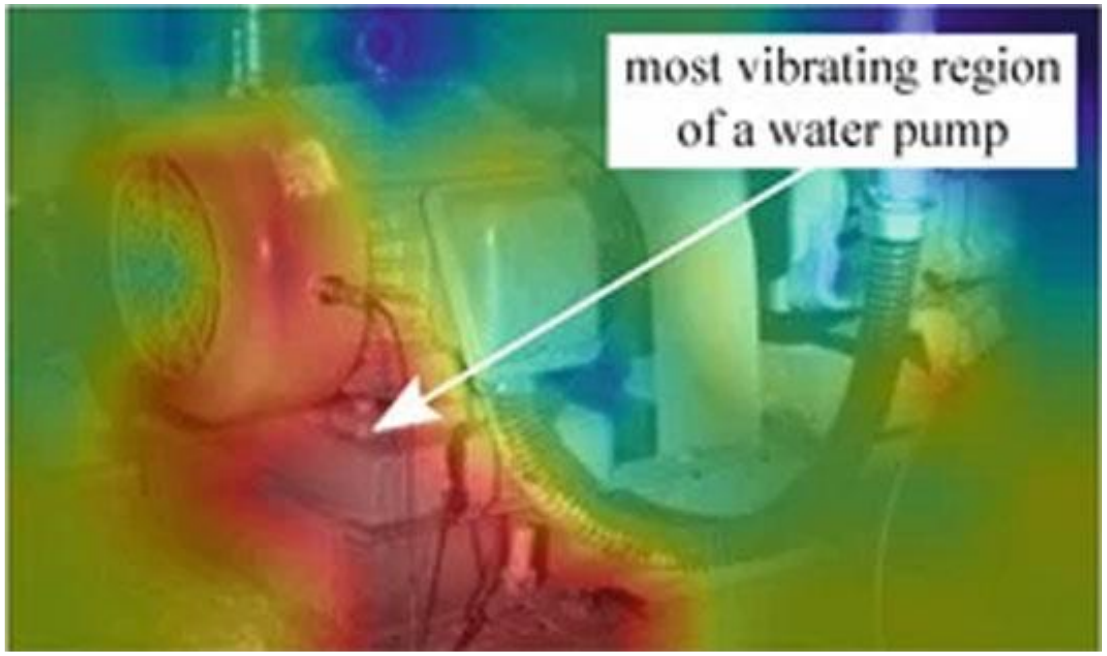
Explanation using confident itemsets explanations:

Minimum confidence threshold: 0.6

Class: LOC: other **Score:** 2.554		**Class:** NUM: count **Score:** 0.666	
Itemset	**Confidence**	**Itemset**	**Confidence**
<where>	0.888	<mile>	0.666
<stadium>	0.666		
<where>, <stadium>	1.0		

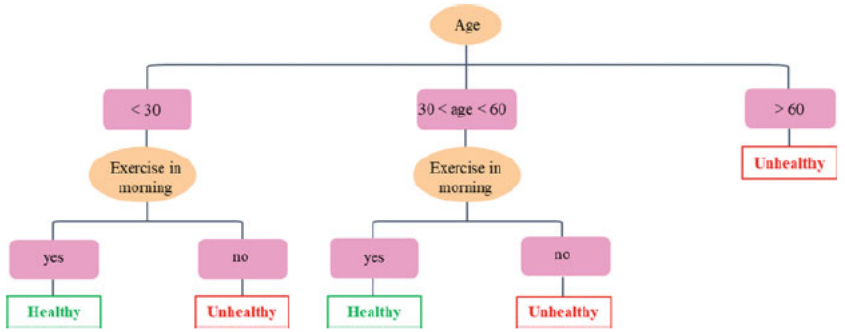
Example of visual explanation

- Visual explanations are one of the most used explanation forms for computer vision tasks.
- Visual explanations use visualization techniques such as class activation maps, gradient-based class activation maps and attention maps to explain the model's prediction.
- Visual explanations are post-hoc methods used for both the local and global explanations.
- Naive users can easily interpret visual explanations, which contain charts, trend lines, etc.



Example of rule-based explanation

- The rule-based explanations are the simplest form of explanation.
- They describe the model inference mechanism using a set of IF–THEN rules or a tree.
- Most rule-based explanations are ante-hoc methods that interpret models with a global scope.
- Ensemble learning and decision tree are popular examples of rule-based explanations.
- This type of explanation is commonly used to develop recommendation systems for naive users.



IF-THEN rules

Is a person healthy?

IF age < 30 AND Exercise in morning = yes, THEN Healthy.

IF age < 30 AND Exercise in morning = no, THEN Unhealthy.

IF 30 < age < 60 AND Exercise in morning = yes, THEN Healthy.

IF 30 < age < 60 AND Exercise in morning = no, THEN Unhealthy.

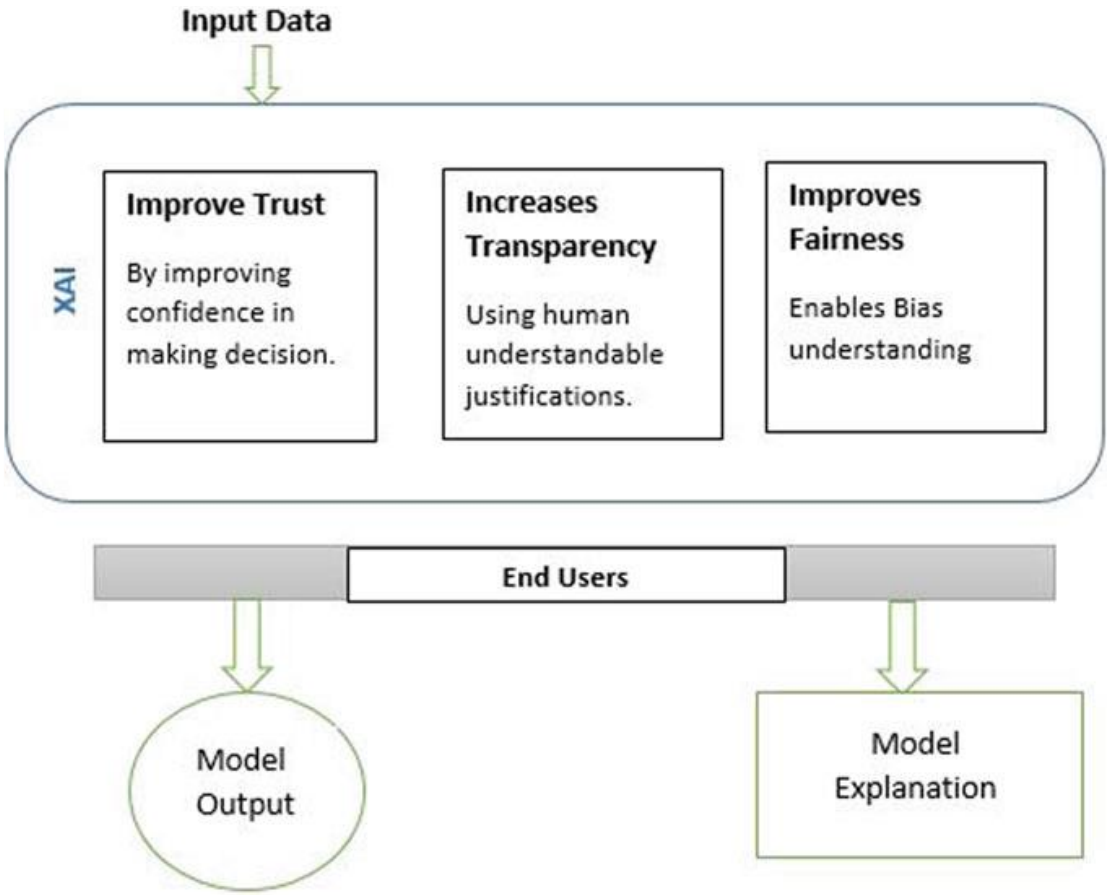
IF age > 60, THEN Unhealthy.

Example of textual explanation

- The textual explanations present the model prediction by learning text explanations as sets of words denoting the features that influence the model prediction.
- They use Natural Language Processing (NLP) techniques to describe the model prediction in natural language.
- Textual explanations are the least common among all forms of explanations due to their high computational requirement for NLP tasks.
- They are suitable for the general user.

Example of textual explanation: *"The person is classified as 'unhealthy' RATHER THAN 'healthy' because person age is more than 30 and no exercise in morning".*

Factors making XAI important



Equivalent key terminology used for XAI

Key term	Description
Black-box AI	The AI model is a black box when it does not reveal anything about the internal design and structure of the system. Because it is difficult for black box models to provide suitable explanations, the problem related to these systems is known as the *black-box problem*.
Interpretable AI	Interpretable systems are those where the users cannot only visualize the parameters necessary for any prediction but can also understand how the input variable is mathematically connected to the outputs. Researchers use the terms interpretability and explainability interchangeably. Others used terms such as comprehensibility or understandability to refer to the same issue, whereas the term interpretable AI is more preferred in the industry.
Responsible AI	This term of XAI takes societal, moral, and ethical values into consideration. The pillars of Responsible AI are Accountability, Transparency, and Responsibility.
Third-wave AI	Recently, the term third-wave in AI has also surfaced, where the system constructs an explanatory model for classifying areal phenomenon and provides reason to their tasks and situations.

Motivation: Why Model Understanding?

Patient Data

25, Female, Cold 32, Male, No 31, Male, Cough .

Predictive Model

Model Understanding

If gender = female,
  if ID_num > 200, then sick

If gender = male,
  if cold = true and cough = true, then sick

Predictions

Healthy Sick Sick . . Healthy Healthy Sick

AUTHORITY

This model is using irrelevant features when predicting on female subpopulation. We cannot approve this!

Motivation: Why Model Understanding?

Utility

Debugging

Bias Detection

Recourse

If and when to trust model predictions

Vet models to assess suitability for deployment

Stakeholders

End users (e.g., loan applicants)

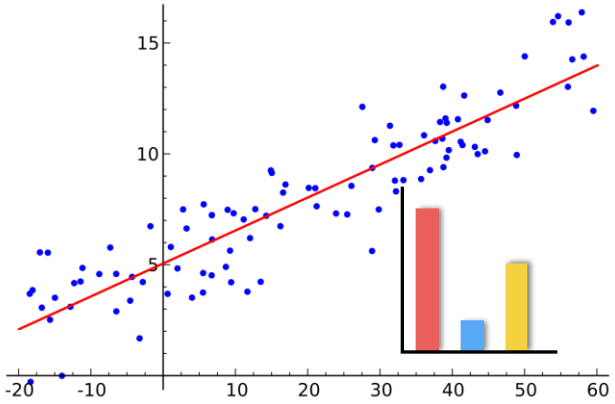
Decision makers (e.g., doctors, judges)

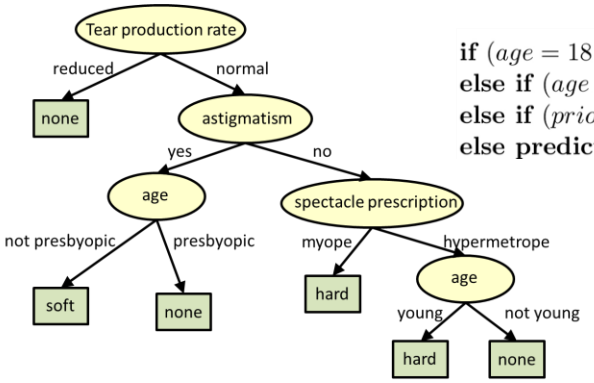
Regulatory agencies (e.g., FDA, European commission)

Researchers and engineers

Achieving Model Understanding

Take 1: Build *inherently interpretable* predictive models

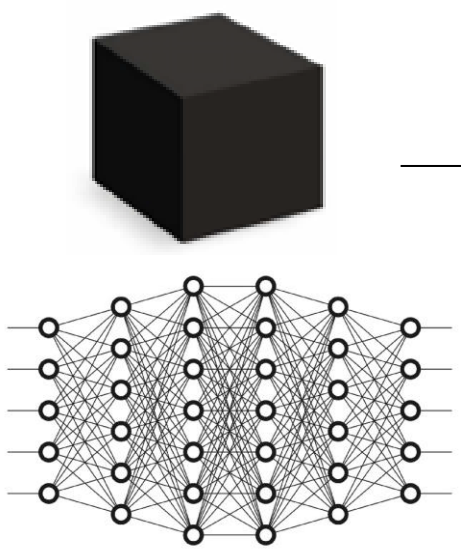


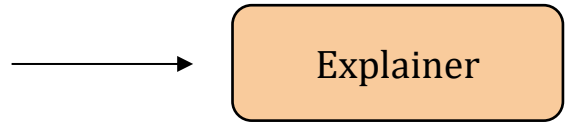


if $(age = 18 - 20)$ **and** $(sex = male)$ **then predict** yes
else if $(age = 21 - 23)$ **and** $(priors = 2 - 3)$ **then predict** yes
else if $(priors > 3)$ **then predict** yes
else predict no

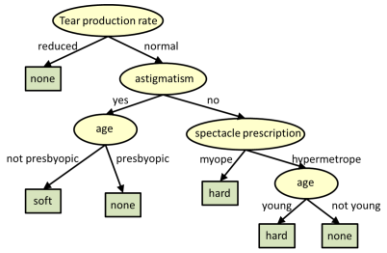
Achieving Model Understanding

Take 2: *Explain* pre-built models *in a post-hoc manner*



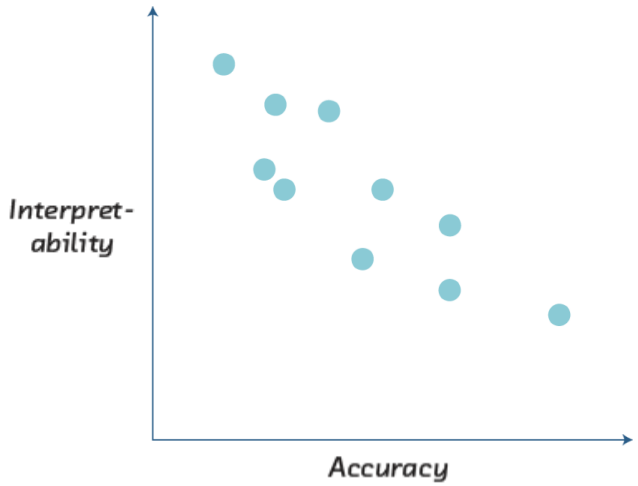


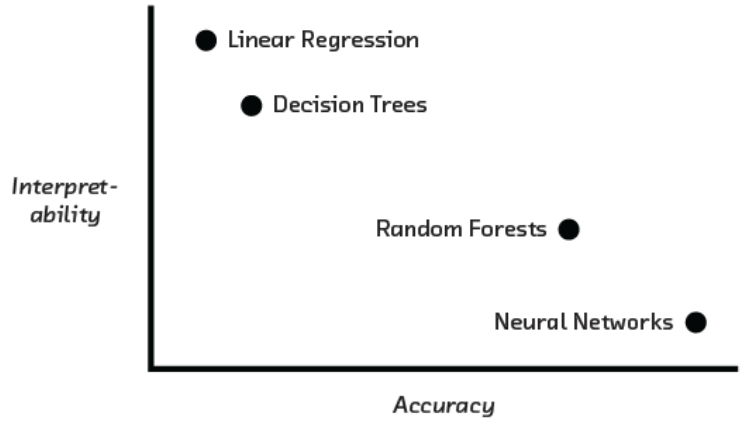
if $(age = 18 - 20)$ **and** $(sex = male)$ **then predict** yes
else if $(age = 21 - 23)$ **and** $(priors = 2 - 3)$ **then predict** yes
else if $(priors > 3)$ **then predict** yes
else predict no



Inherently Interpretable Models vs. Post hoc Explanations

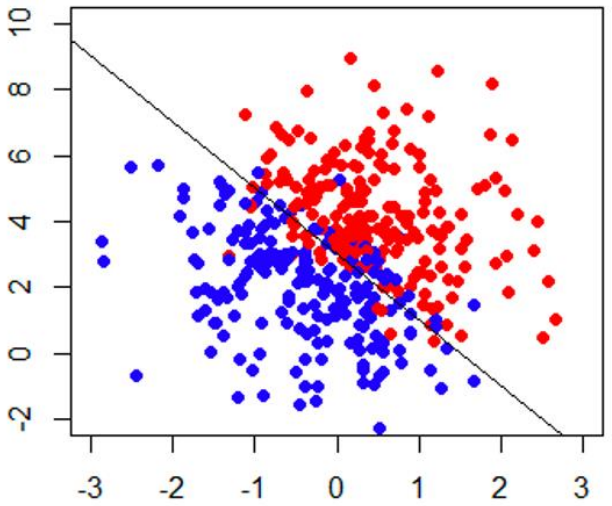
Example



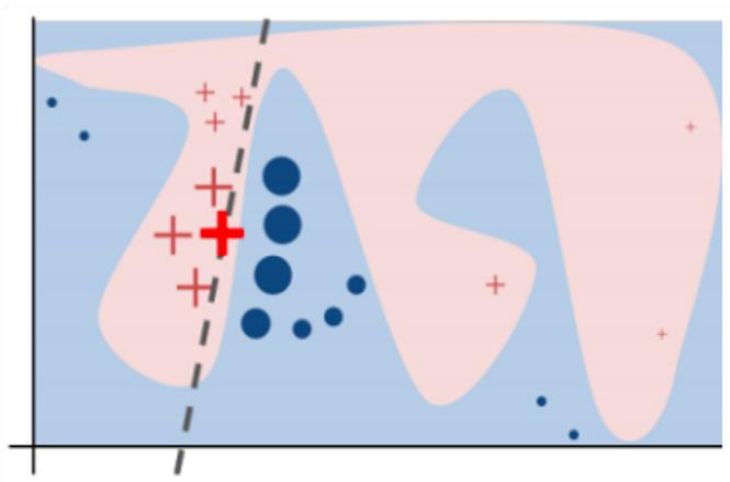


In ***certain*** settings, *accuracy-interpretability trade offs* may exist.

Inherently Interpretable Models vs. Post hoc Explanations



can build interpretable + accurate models



complex models might achieve higher accuracy

Inherently Interpretable Models vs. Post hoc Explanations

Sometimes, you don't have enough data to build your model from scratch.

And all you have is a (proprietary) black box!