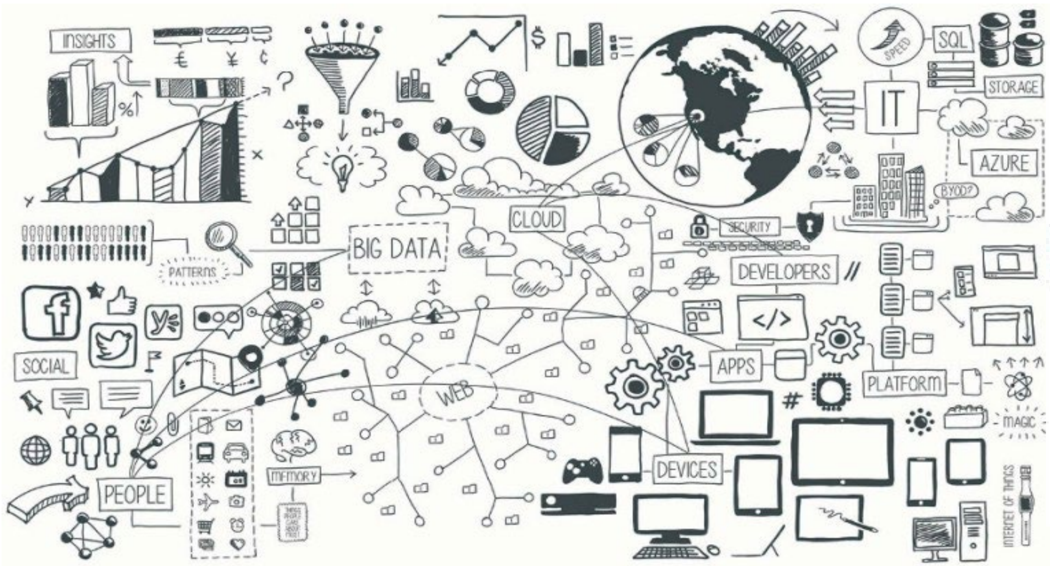
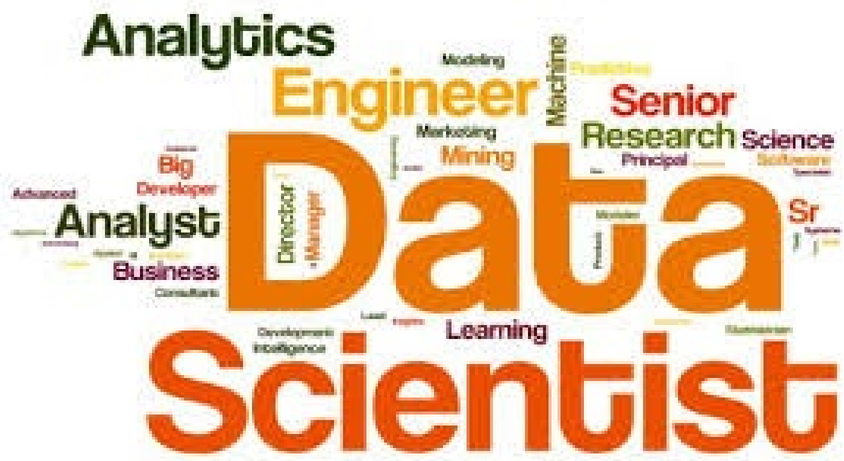
Data Mining (Minería de Datos)

Explainable Artificial Intelligence (XAI)





Sixto Herrera, Ana Casanueva

Univ. de Cantabria - MACC

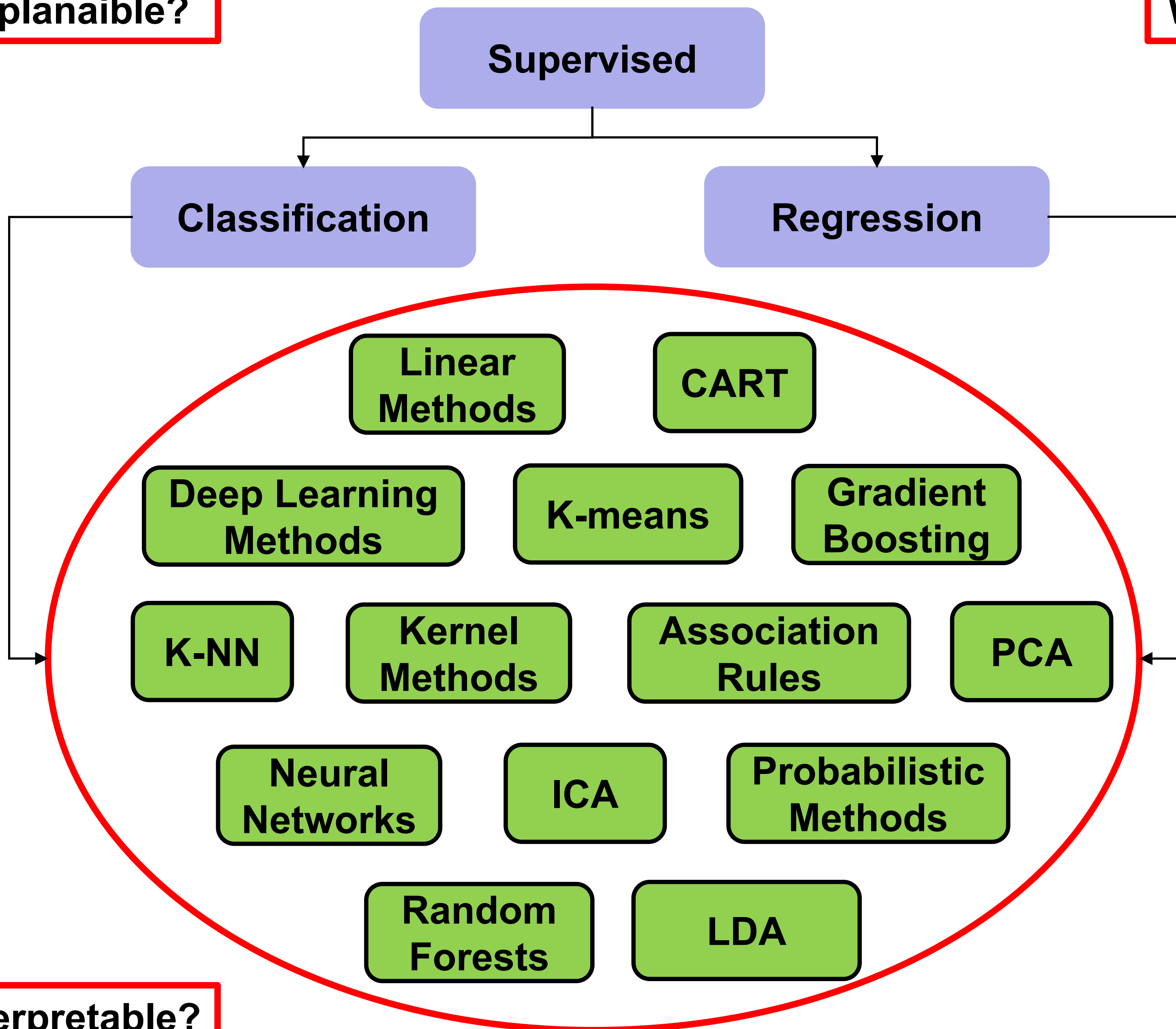
UNIVERSIDAD DE CANTABRIA

M1966 – Data Mining

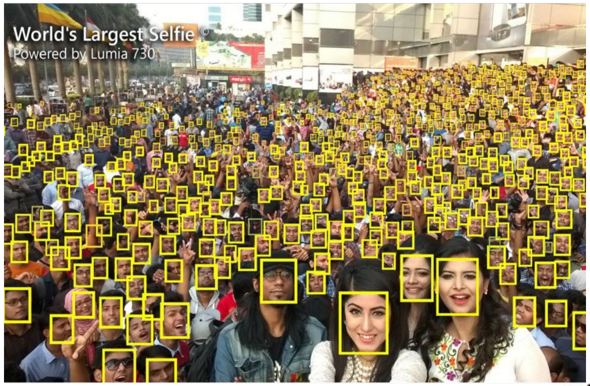
Oct	29	M	Presentación, Introducción y Perspectiva histórica (2h, T)
	30	X	Paradigmas de Aprendizaje, Problemas Canónicos y Datasets (2h, T-L)
	31	J	Reglas de Asociación (2h, T)
Nov	4	L	Reglas de Asociación (2h, L)
	6	X	Evaluación, Sobreajuste y Cross-Validation (2h, T)
	11	L	Cross-Validation (2h, L)
	13	X	Árboles de Clasficación y Decisión (2h, T)
	18	L	Árboles de Clasficación y Decisión (2h, L)
	20	X	Técnicas de Vecinos Cercanos, (k-NN) (2h, T)
	25	L	Técnicas de Vecinos Cercanos, (k-NN) (2h, L)
	27	X	Comparación de Técnicas de Clasificación (2h, L)
Dic	2	L	Árboles de Regresión (CART) (2h, T)
	4	X	Árboles de Regresión (CART) (2h, V, 17:30-19:30)
	9	L	Paquete CARET (2h, L, 17:30-19:30)
	11	X	Ensembles: Bagging and Boosting (2h, T)
	13	V	Random Forests (2h, L)
	16	L	Gradient Boosting (2h, T-L)
	18	X	XAI - Explainable Artificial Intelligence (2h, T-L, 17:30-19:30)
Ene	8	X	Reducción de la Dimensión (No lineal) (2h, T-L)
	13	L	Reducción de la Dimensión (No lineal) (2h, T-L)
	15	X	Técnicas de Agrupamiento (2h, T)
	20	L	Técnicas de Agrupamiento (2h, L)
	22	X	Predicción Condicionada (2h, L)
	24	V	Sesión de Repaso (2h, T-L)
	29	X	Examen//Cuestionario (2h, T-L)

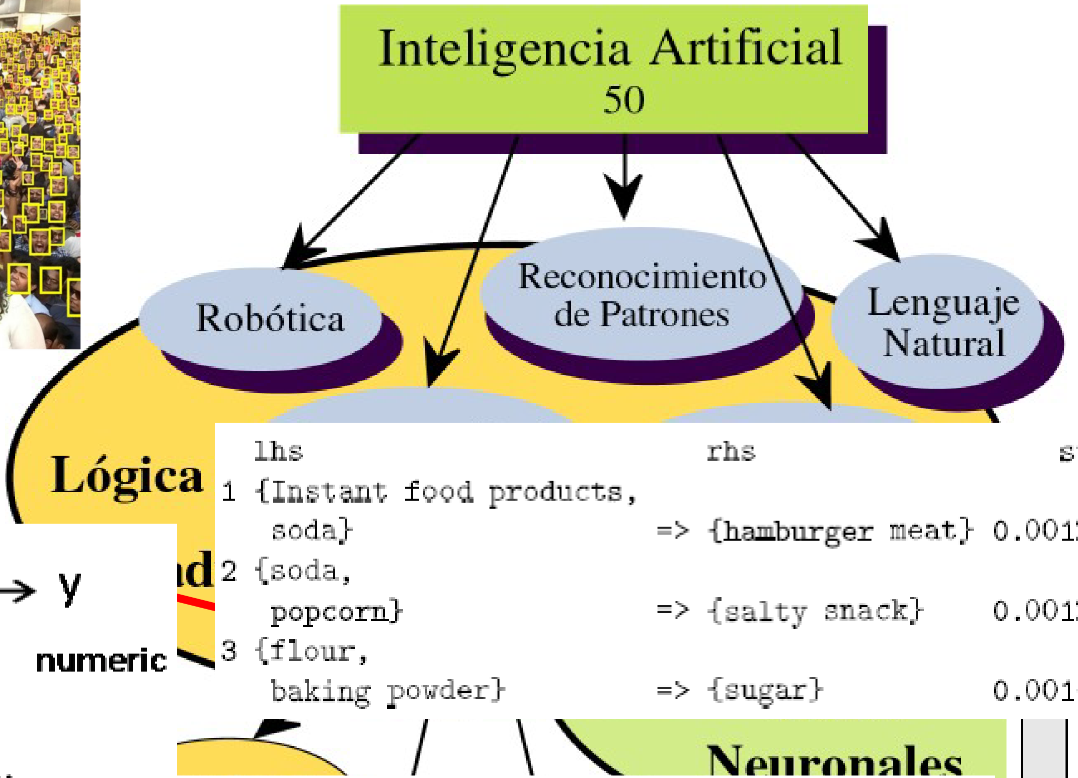
Are they explanaible?

Which ones?



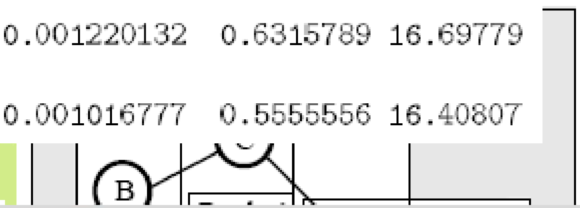
Are they interpretable?



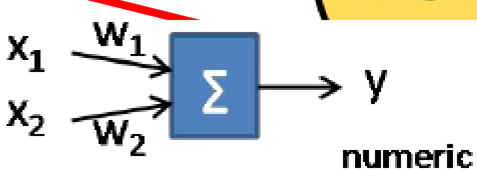


H, I cierto
L falso

	lhs		rhs	support	confidence	lift
1	{Instant food products, soda}	=>	{hamburger meat}	0.001220132	0.6315789	18.99565
2	{soda, popcorn}	=>	{salty snack}	0.001220132	0.6315789	16.69779
3	{flour, baking powder}	=>	{sugar}	0.001016777	0.5555556	16.40807



DATA MINING
1990



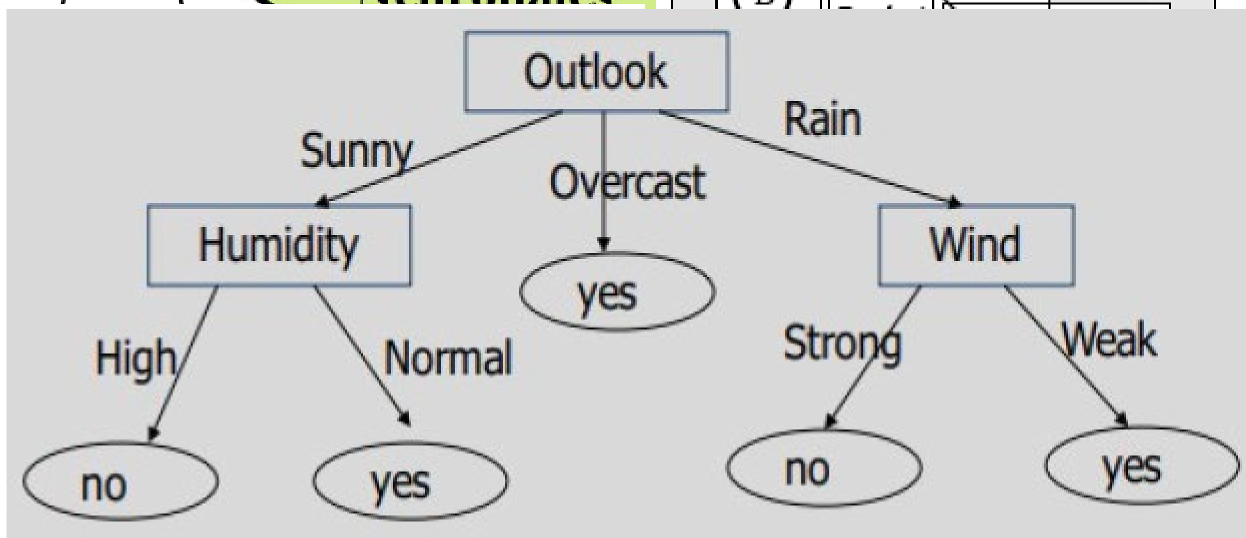
numeric or binary

$$y = w_0 + w_1x_1 + w_2x_2$$

$$y = f(\mathbf{X},\mathbf{W}) = \mathbf{X}^T.\mathbf{W}$$

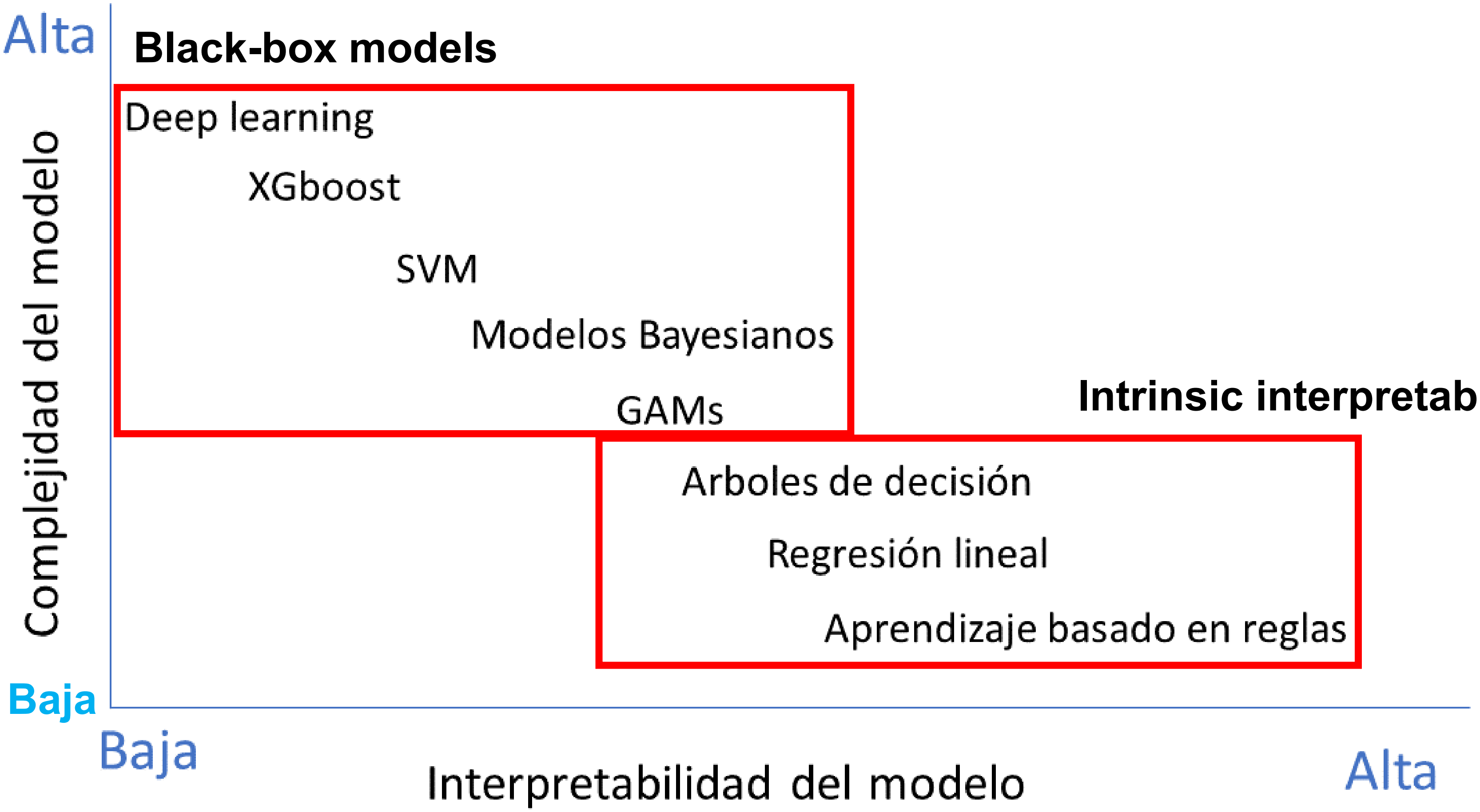
Human-like reasoning

Logical inference, look for relations on graphs, etc.

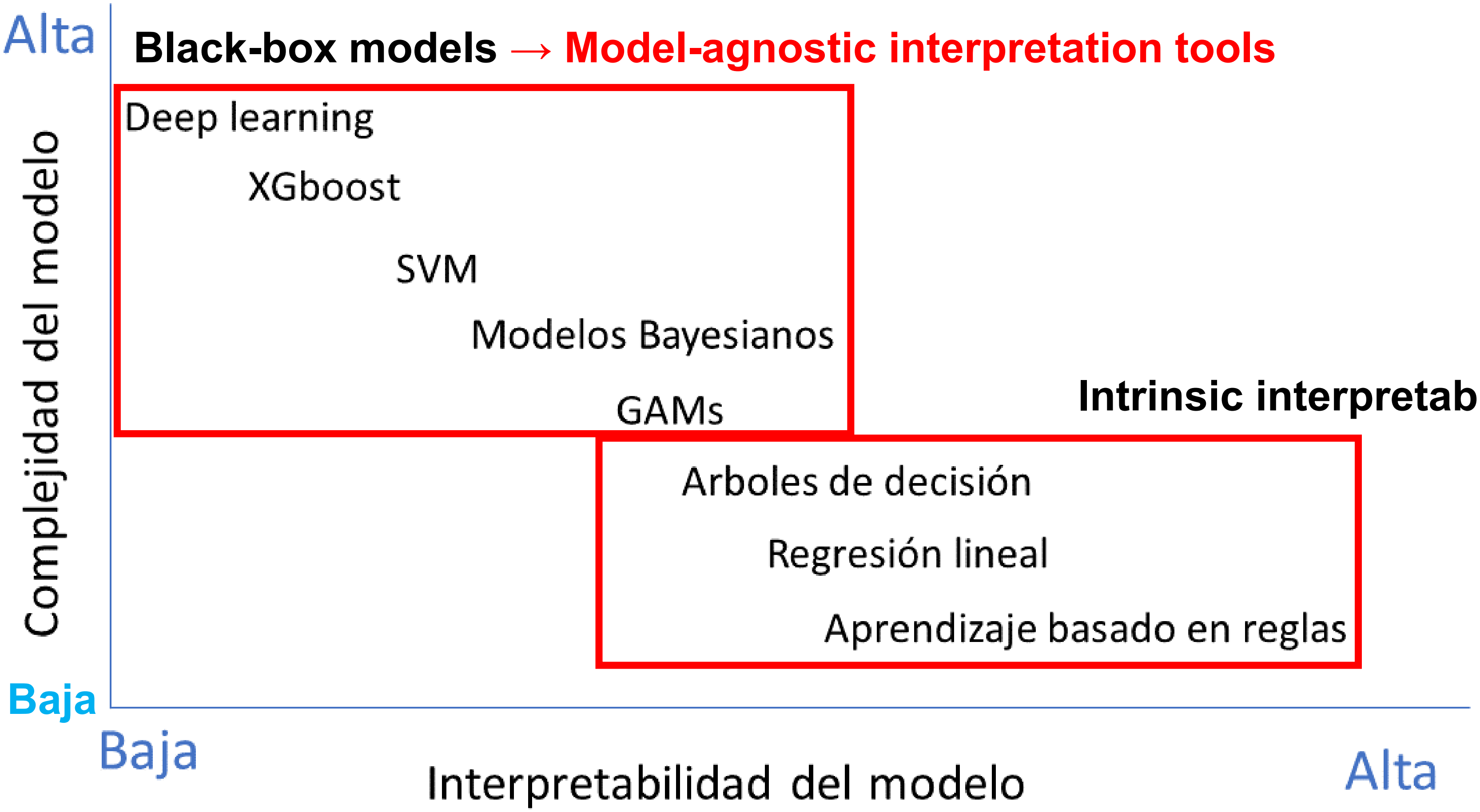


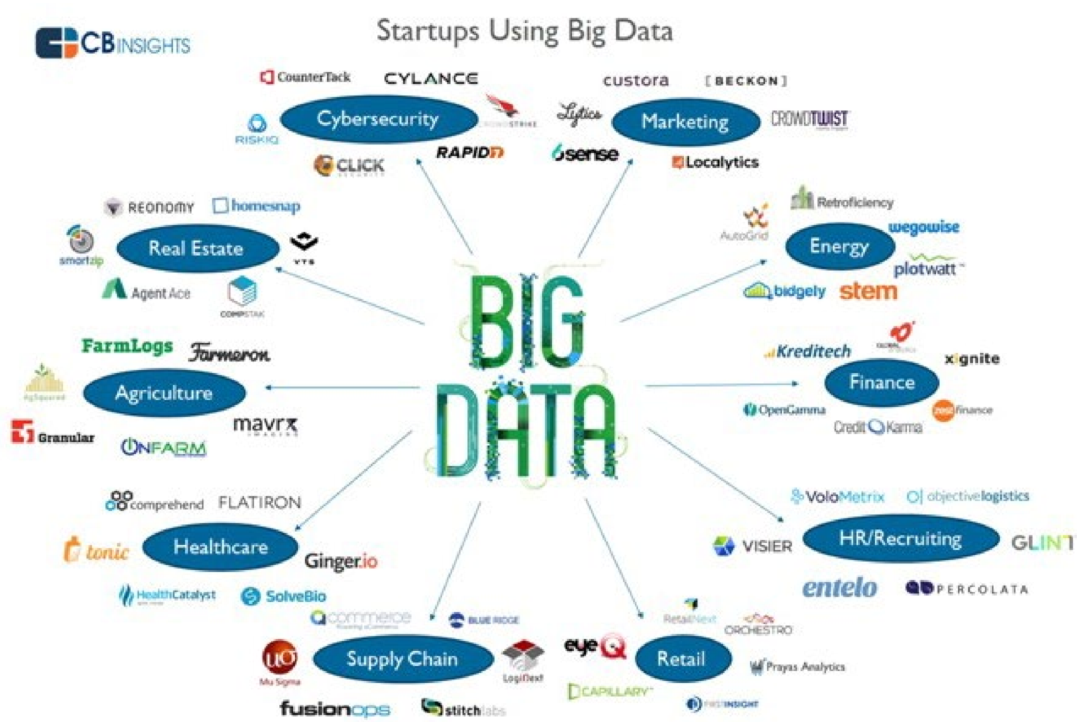
computing power

Are they explanaible? Which ones?



Are they explanaible? Which ones?





Artificial Intelligence (AI) models is used to:

avoid accidents in cars
manage investment in banks
loan decisions in banks
aid doctors diagnosing in hospitals
detect diseases in hospitals
help officials recover evidence in law enforcement
make law enforcement easier
military purposes of many countries
risk-detection in insurance organizations
…

Startups Using Big Data

For a model to be embraced by end-users and industries, it must be trustworthy (fairness, robustness, interpretability, and explainability/interpretation).

Requerimientos regulatorios
Falta de confianza
Potencial mal uso
Impactos sociales y humanos
….

Artificial Intelligence (AI) models is used to:

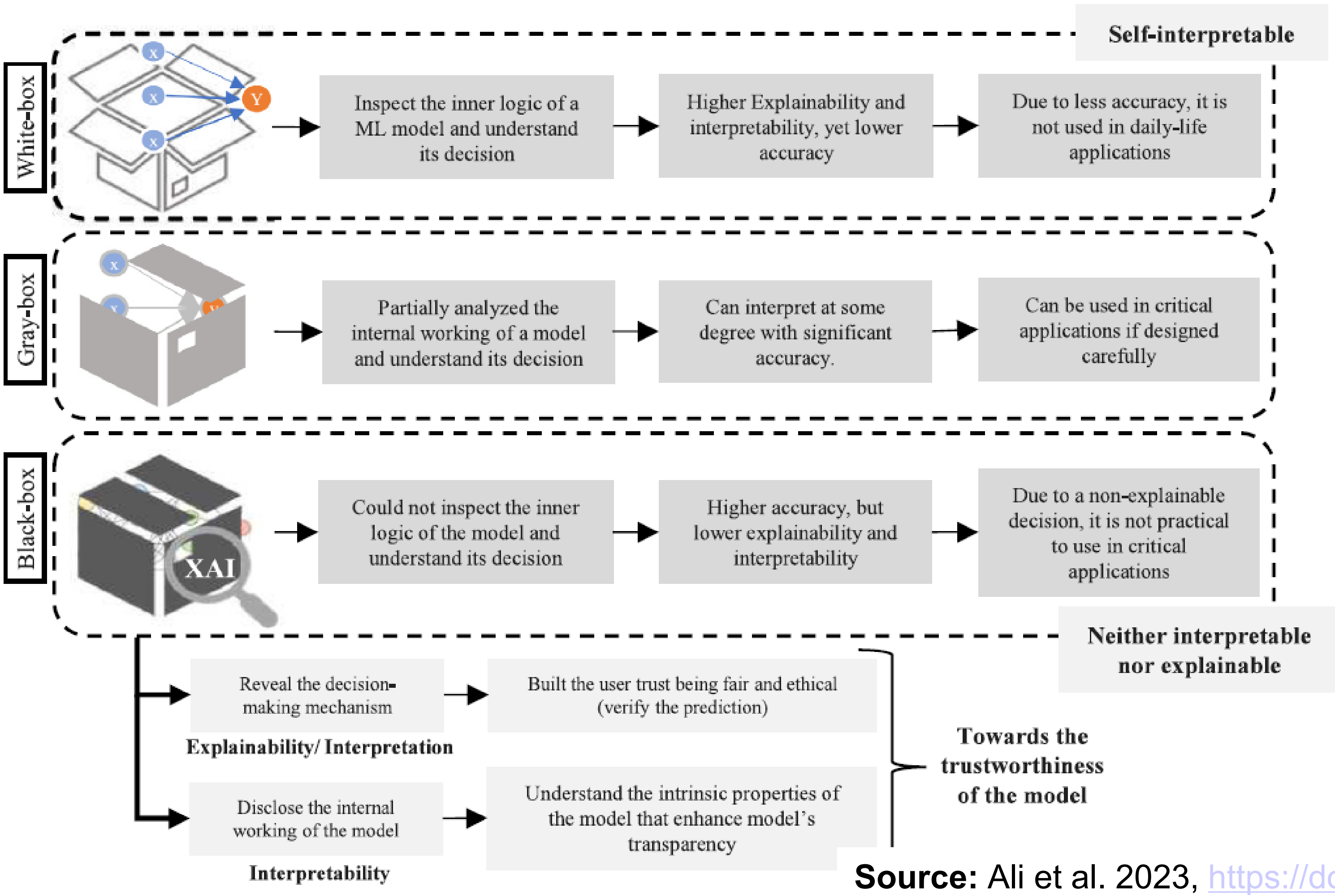
Avoid accidents in cars
Manage investment in banks
Loan decisions in banks
Aid doctors diagnosing in hospitals
Detect diseases in hospitals
Help officials recover evidence in law enforcement

Código de Ética y Conducta Profesional de ACM:
https://www.acm.org/code-of-ethics/the-code-in-spanish
https://www.unesco.org/en/artificial-intelligence/recommendation-ethics

What do interpretability and explainability mean?

Explainability provides insight into the AI-model decision to the **end-user** in order to **build trust** that the AI is making **correct** and **non-biased** decisions based on facts.

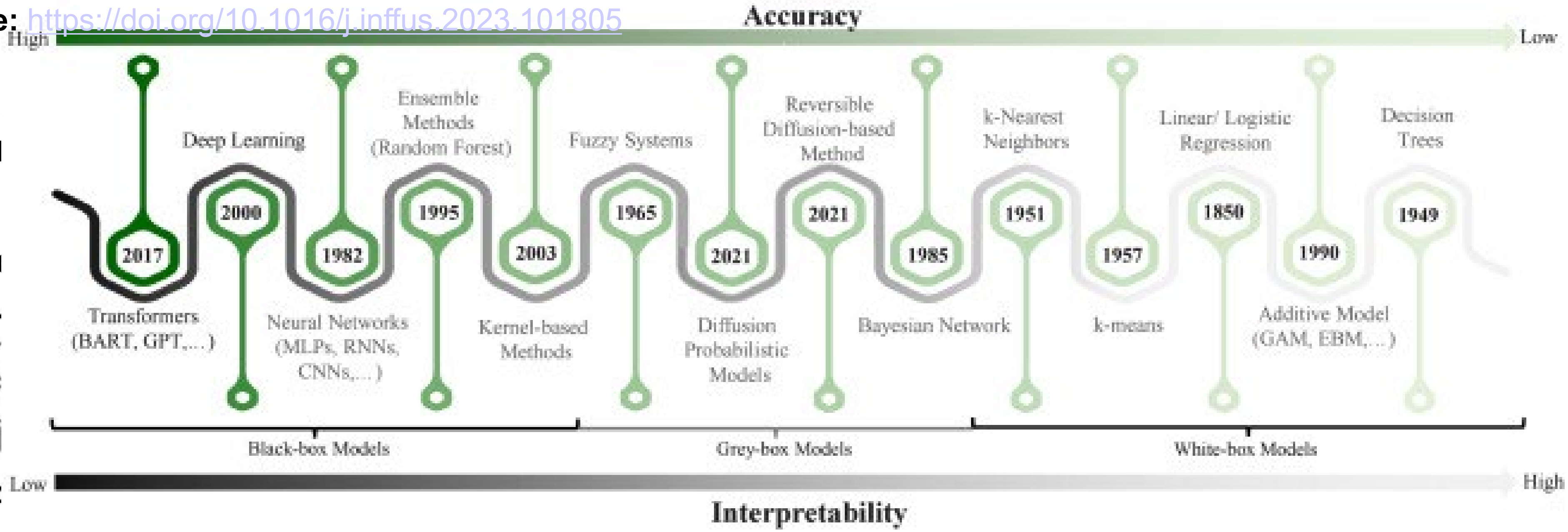
Interpretability enables **developers** to delve into the model's decision making process, boosting their confidence in understanding where the model gets its results.

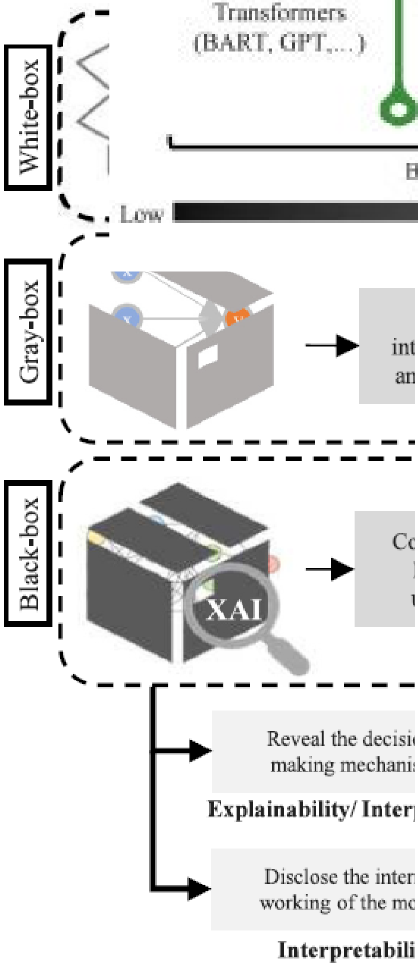


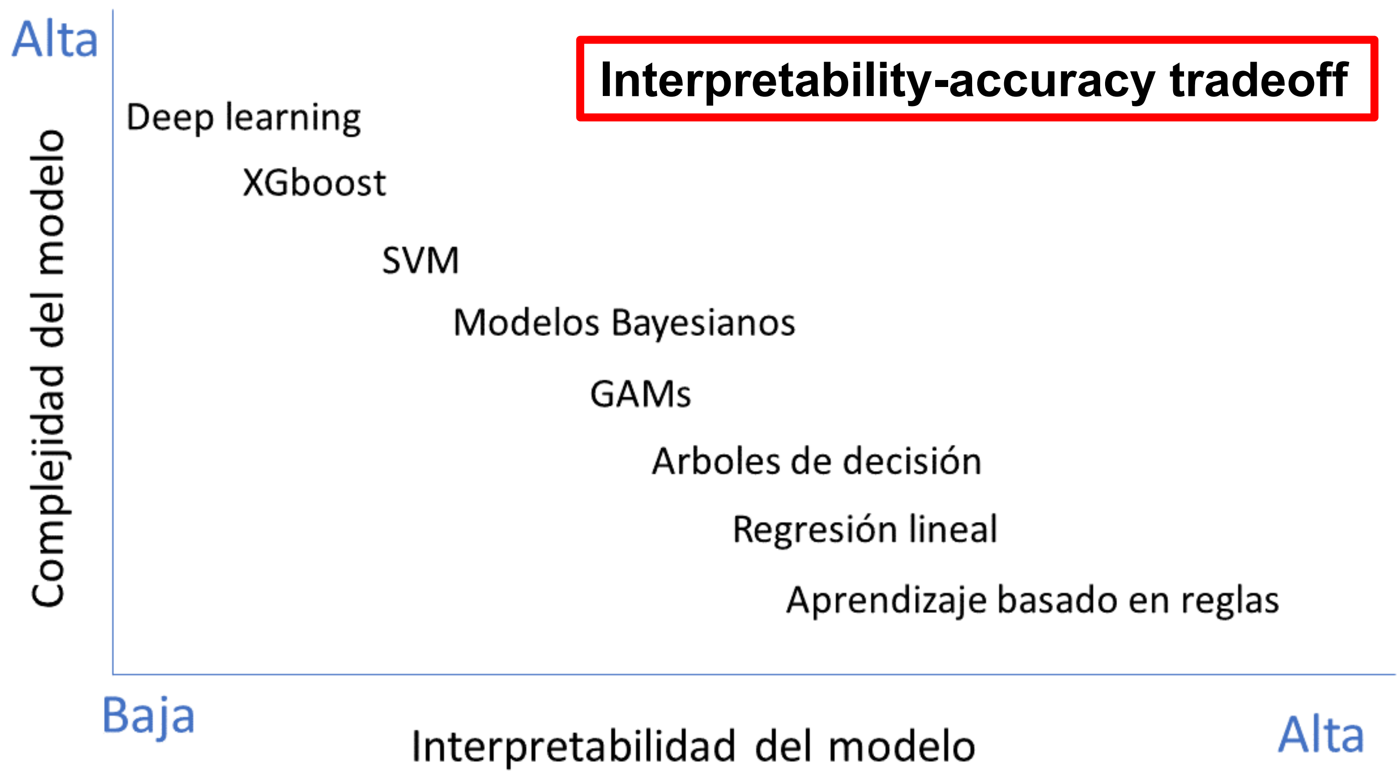
Source: Ali et al. 2023, https://doi.org/10.1016/j.inffus.2023.101805

Source: https://doi.org/10.1016/j.inffus.2023.101805

Ex
tru
Int
bo
n?







For a model to be embraced by end-users and industries, it must be trustworthy (fairness, robustness, interpretability, and explainability/interpretation).

eXplainable Artificial Intelligence (XAI) techniques are aimed at producing ML models with a good **interpretability-accuracy tradeoff** via:

(i) building white/gray-box ML models which are interpretable by design (at least at some degree) while achieving high accuracy

(ii) endowing black-box models with a minimum level of interpretability when white/gray-box models are not able to achieve an admissible level of accuracy.

The goal of XAI research is to make AI systems more comprehensible and transparent to humans without sacrificing performance. The primary goal of XAI is to obtain human-interpretable models, but there are others:

- To empower individuals to combat any negative consequences of automated decision-making.
- To assist individuals in making more informed choices.
- To expose and protect security vulnerabilities.
- To integrate algorithms with human values is an important goal.
- To enhance industry standards for the development of AI-powered products, thus improving consumer and business confidence.
- To enforce the Right of Explanation policy.

Interpretability enables **developers** to delve into the model's decision making process, boosting their confidence in understanding where the model gets its results.

For a model to be embraced by end-users and industries, it must be trustworthy (fairness, robustness, interpretability, and explainability/interpretation).

The more a machine's decision affects a person's life, the more important it is for the machine to explain its behavior. AI Act of the European Union.

When we do not need interpretability: the model has no significant impact and/or the problem is well studied. On the other hand, Interpretability might enable people or programs to manipulate the system.

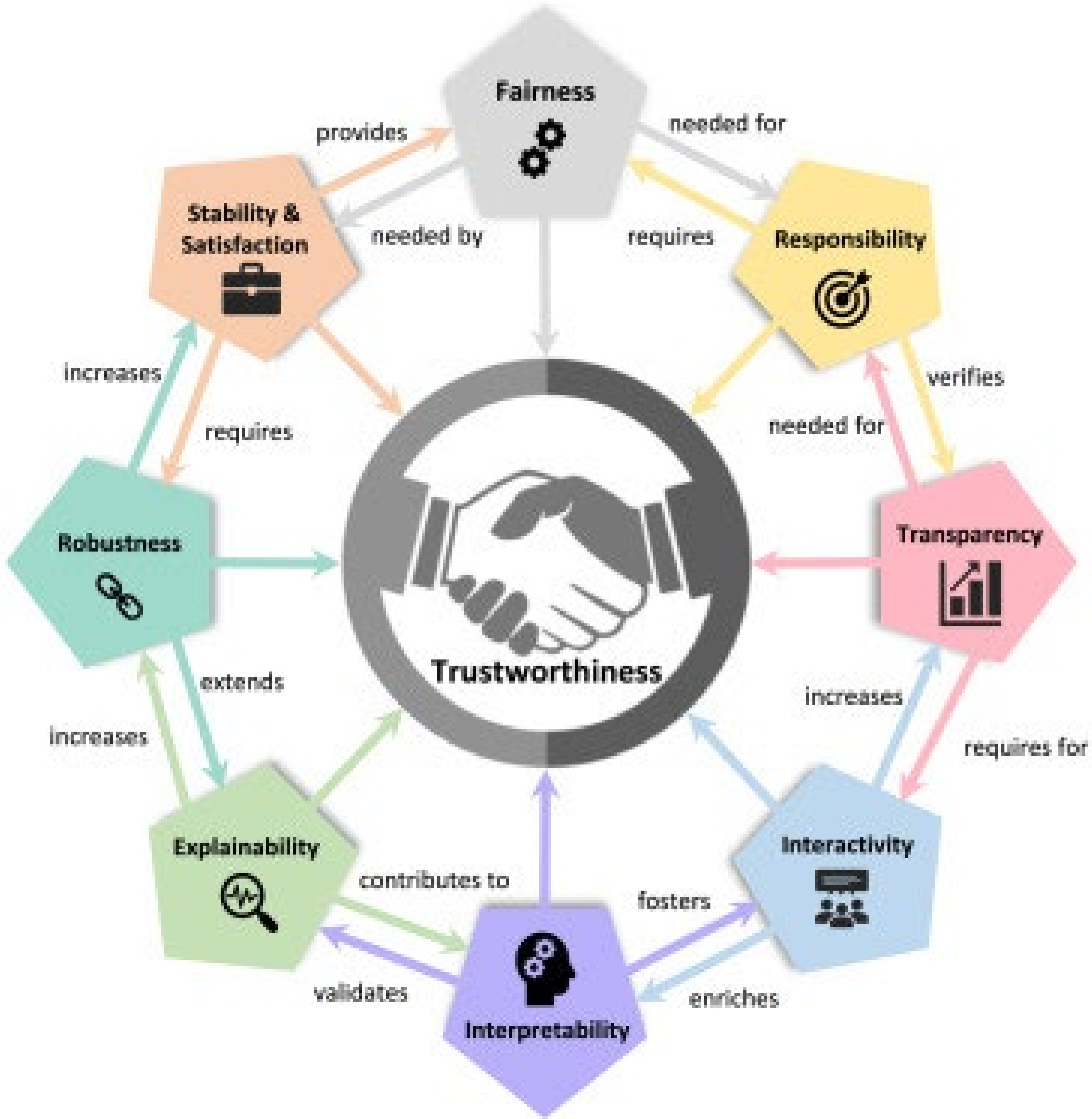
Machine learning models can only be debugged and audited when they can be interpreted. If you can ensure that the machine learning model can explain decisions, you can also check the following traits more easily (Doshi-Velez and Kim 2017):

- **Fairness**: Ensuring that predictions are unbiased and do not implicitly or explicitly discriminate against underrepresented groups.
- **Privacy**: Ensuring that sensitive information in the data is protected.
- **Reliability or Robustness**: Ensuring that small changes in the input do not lead to large changes in the prediction.
- **Causality**: Check that only causal relationships are picked up.
- **Trust**: It is easier for humans to trust a system that explains its decisions compared to a black box.

For a model to be embraced by end-users and industries, it must be trustworthy (fairness, robustness, interpretability, and explainability/interpretation).

The more a machine's decision affects a person's life, the more important it is for the machine to explain its behavior. AI Act of the European Union.

When we do not ne ct and/or the problem is well studied. On th rograms to manipulate the syste



Machine learning n can be interpreted. If you can ensure t , you can also check the following traits

- **Fairness**: Ensu or explicitly discriminate aga
- **Privacy**: Ensuri
- **Reliability or R** do not lead to large changes in the p
- **Causality**: Che
- **Trust**: It is easie ons compared to a black box.

Machine Learning Interpretability (https://christophm.github.io/interpretable-ml-book/)

1) Use interpretable models, such as linear models or decision trees
2) Use of model-agnostic interpretation tools that can be applied to any supervised machine learning model. Model-agnostic methods can be divided into:
   1) global methods that describe the average behavior of the model
   2) local methods that explain individual predictions

Do you just want to know what is predicted? A diagnostic and treatment

Or do you want to know why the prediction was made? If it fails, why has it failed? Knowing the 'why' can help you learn more about the problem, the data and the reason why a model might fail.

Based on the AI-system of the bank I can not get a loan, why? It is an objective reason or the AI-system could have any learned demographic (e.g. racial) bias.

Local or global? Does the interpretation method explain an individual prediction or the entire model behavior?
Scope of Interpretability:
Algorithm Transparency: How does the algorithm create the model?
Global, Holistic Model Interpretability: How does the trained model make predictions?
Global Model Interpretability on a Modular Level: How do parts of the model affect predictions?
Local Interpretability for a Single Prediction: Why did the model make a certain prediction for an instance? Local explanations can therefore be more accurate than global explanations.
Local Interpretability for a Group of Predictions: Why did the model make specific predictions for a group of instances?

Machine Learning Interpretability (https://christophm.github.io/interpretable-ml-book/)

1) Use interpretable models, such as linear models or decision trees
2) Use of model-agnostic interpretation tools that can be applied to any supervised machine learning model. Model-agnostic methods can be divided into:
    1) global methods that describe the average behavior of the model
    2) local methods that explain individual predictions

Example-Based Explanations

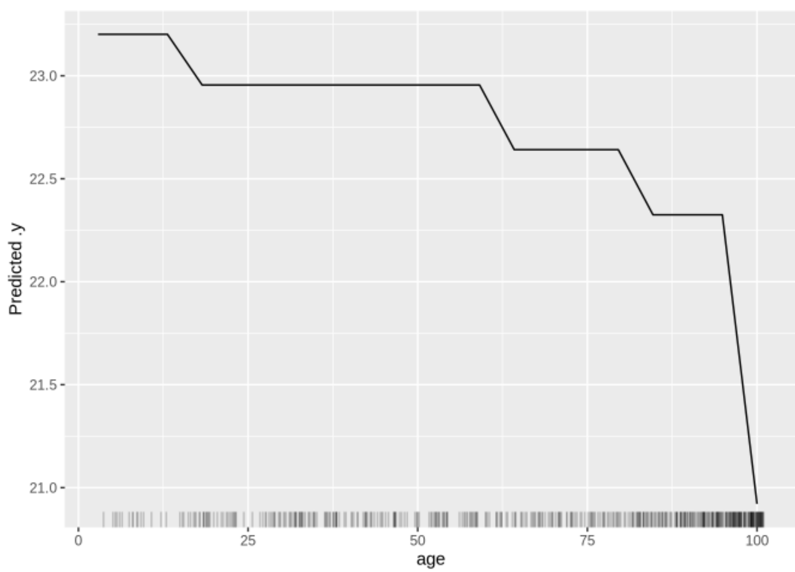
Example-based explanation methods select particular instances of the dataset to explain the behavior of machine learning models or to explain the underlying data distribution. They are mostly model-agnostic, because they make any machine learning model more interpretable.

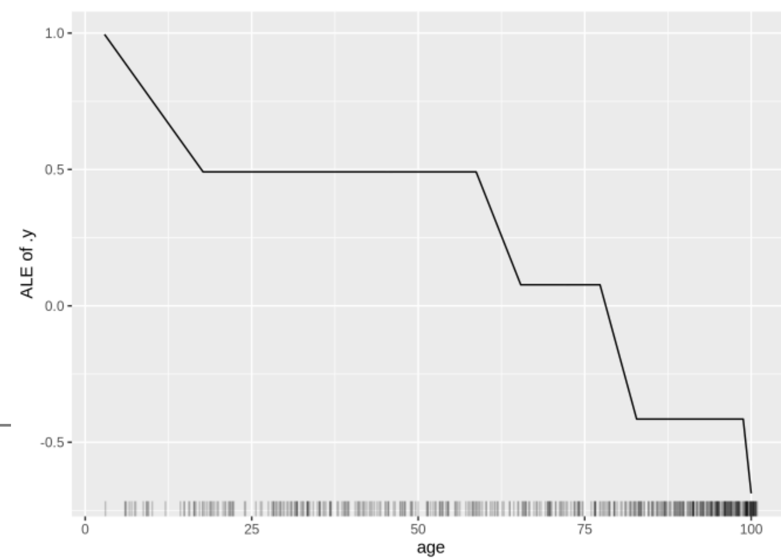
Global Model-Agnostic Methods:

- Feature effect plots: PDP and ALE.

$$\begin{aligned}\hat{f}_{S,PDP}(x) &= E_{X_C}\left[\hat{f}(x_S, X_C)\right] \\ &= \int_{X_C} \hat{f}(x_S, X_C)\, d\mathbb{P}(X_C)\end{aligned}$$

$$\begin{aligned}\hat{f}_{S,ALE}(x_S) &= \int_{z_{0,S}}^{x_S} E_{X_C|X_S=x_S}\left[\hat{f}^S(X_s, X_c)|X_S = z_S\right] dz_S - \text{constant} \\ &= \int_{z_{0,S}}^{x_S} \left(\int_{x_C} \hat{f}^S(z_s, X_c)\, d\mathbb{P}(X_C|X_S = z_S)\, d\right) dz_S - \text{constant}\end{aligned}$$





Machine Learning Interpretability (https://christophm.github.io/interpretable-ml-book/)

1) Use interpretable models, such as linear models or decision trees
2) Use of model-agnostic interpretation tools that can be applied to any supervised machine learning model. Model-agnostic methods can be divided into:
   1) global methods that describe the average behavior of the model
   2) local methods that explain individual predictions

Example-Based Explanations

Example-based explanation methods select particular instances of the dataset to explain the behavior of machine learning models or to explain the underlying data distribution. They are mostly model-agnostic, because they make any machine learning model more interpretable.

Global Model-Agnostic Methods:

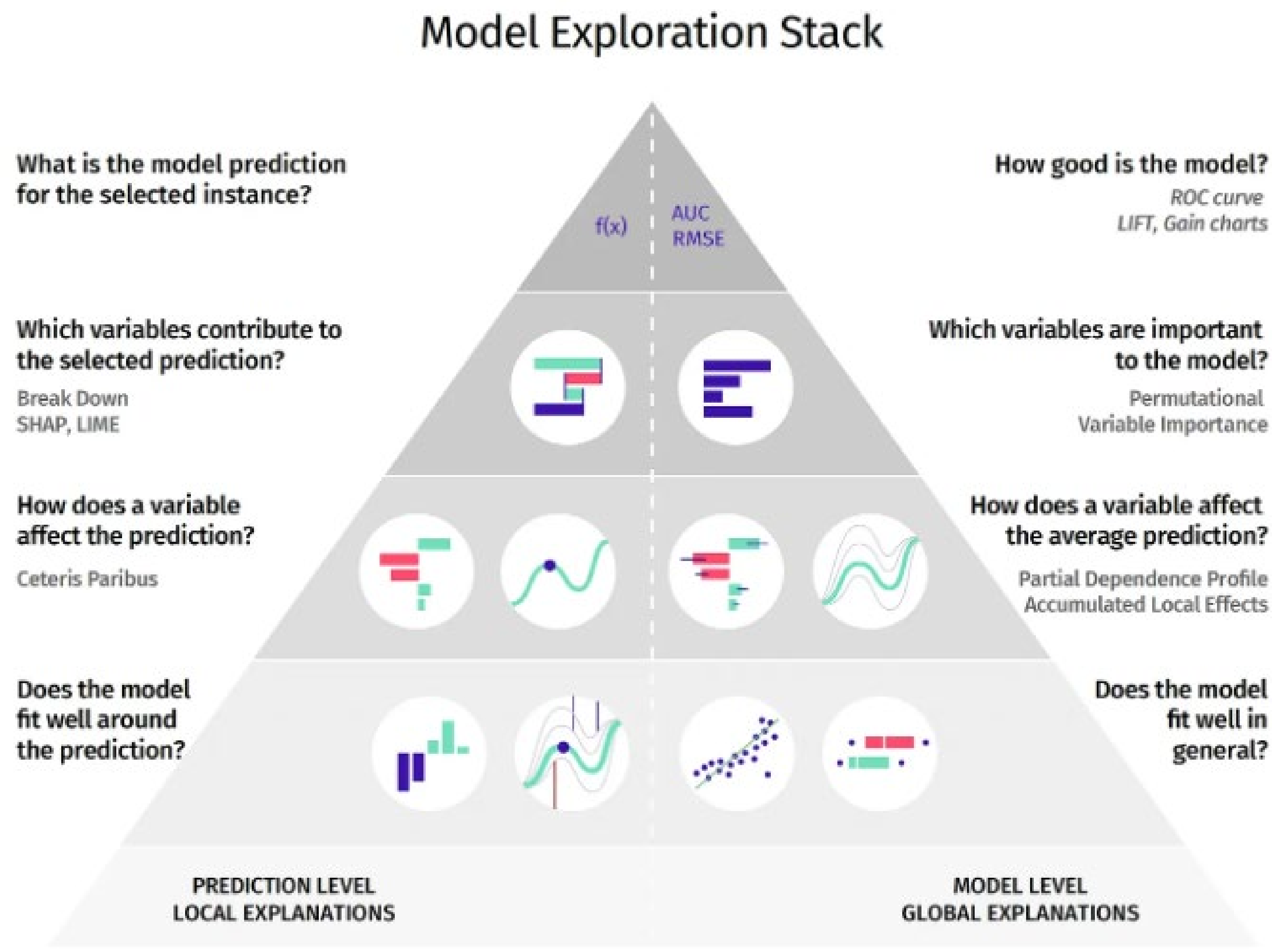
- Feature effect plots: PDP and ALE.
- Feature importance and interactions (H-statistic) quantifies to what extent the prediction is the result of joint effects of the features.
- Functional decomposition is a central idea of interpretability and a technique that decomposes the complex prediction function into smaller parts.
- Global surrogate models replaces the original model with a simpler model for interpretation.
- Prototypes and criticisms are representative data point of a distribution and can be used to enhance interpretability.

Local Model-Agnostic Methods:

- The Individual Conditional Expectation (ICE) plot.
- Local surrogate models (LIME).
- Counterfactual Explanations.
- …

Machine Learning Interpretability (https://christophm.github.io/interpretable-ml-book/)

1) Use interpr
2) Use of mod   achine learning
   model. Moc
   1) global
   2) local n



Biecek, P. and Burzykowski, T. Explanatory Model Analysis

Example-Ba

Example-basec   in the behavior
of machine lea   stly model-
agnostic, becau

Global Model-

- Feature effe
- Feature imp   ction is the result
  of joint effec
- Functional c   composes the
  complex pre
- Global surro   retation.
- Prototypes a   used to enhance
  interpretabil

Local Model-A

- The Individu
- Local surrog
- Counterfactu
- …

Source: https://medium.com/responsibleml/basic-xai-with-dalex-part-1-introduction-e68f65fa2889