Merci à nos sponsors

THE INTERNATIONAL INSTITUTE OF
SUPINFO
INFORMATION TECHNOLOGY

nextdecision

AZEO
talents & technology

avanade

Et en partenariat avec le meetup Azure Nantes

Nantes
Microsoft
Azure

XAI : eXplainable Artificial Intelligence

Explainability and AI - Definitions

Explainability and AI - Definitions

Artificial Intelligence :

“[Artificial Intelligence] refers to a programme whose ambitious objective is to understand and reproduce human cognition; creating cognitive processes comparable to those found in human beings” for a meaningful artificial intelligence towards a french and european strategy,Villani report, 2018

Explainability and AI - Definitions

XAI : eXplainable Artificial Intelligence

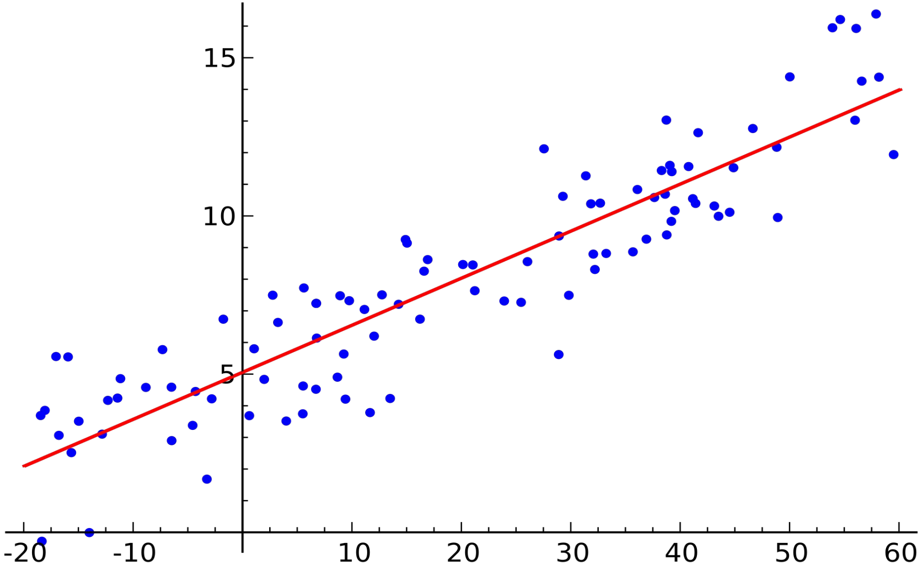
Explainability :

Make the algorithms inner state understable by humans

Explainability and AI - An exponential complexity growth

Linear regression

Two correlated axes

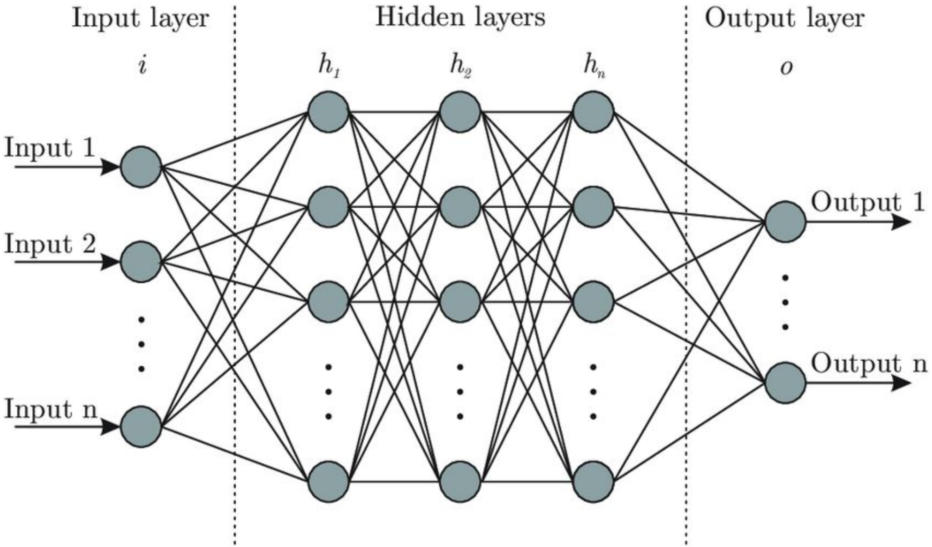


Wikipedia: Linear regression article

Explainability and AI - An exponential complexity growth

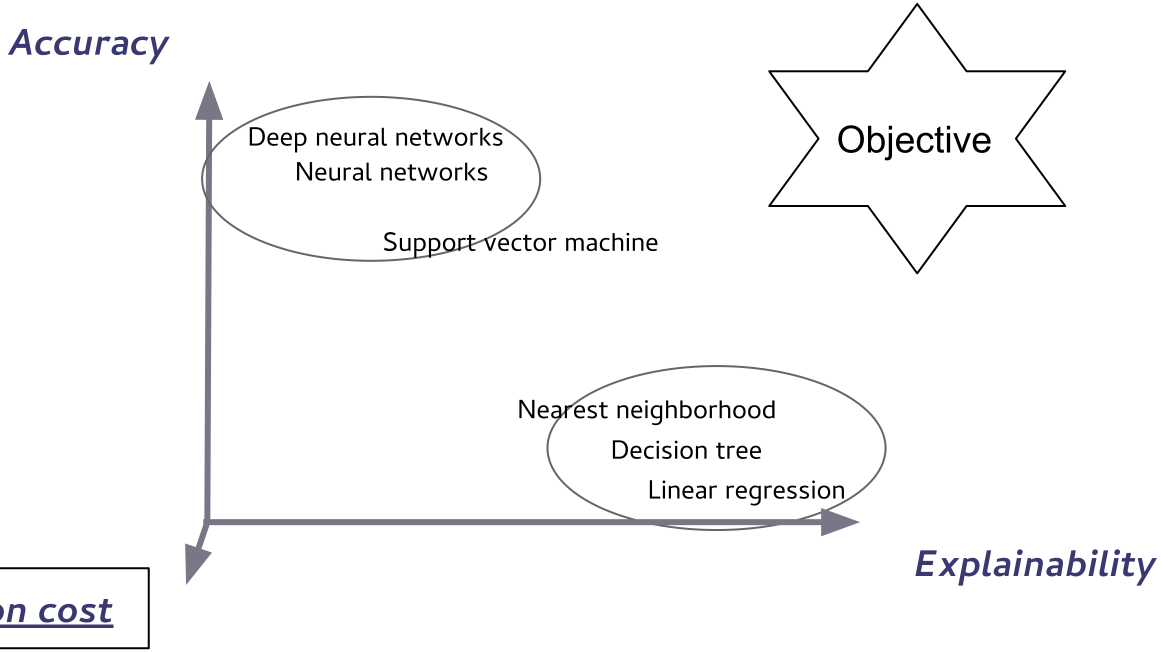
Artificial neural network

Many fields of low value, very entangled



Bre, Facundo & Gimenez, Juan & D. Fachinotti, Víctor. (2017). Prediction of wind pressure coefficients on building surfaces using Artificial Neural Networks. Energy and Buildings.

Explainability and AI - Accuracy VS explainability



Why is understanding the black box important ?

Why is understanding the black box important ?

- General Data Protection Regulation (GDPR)

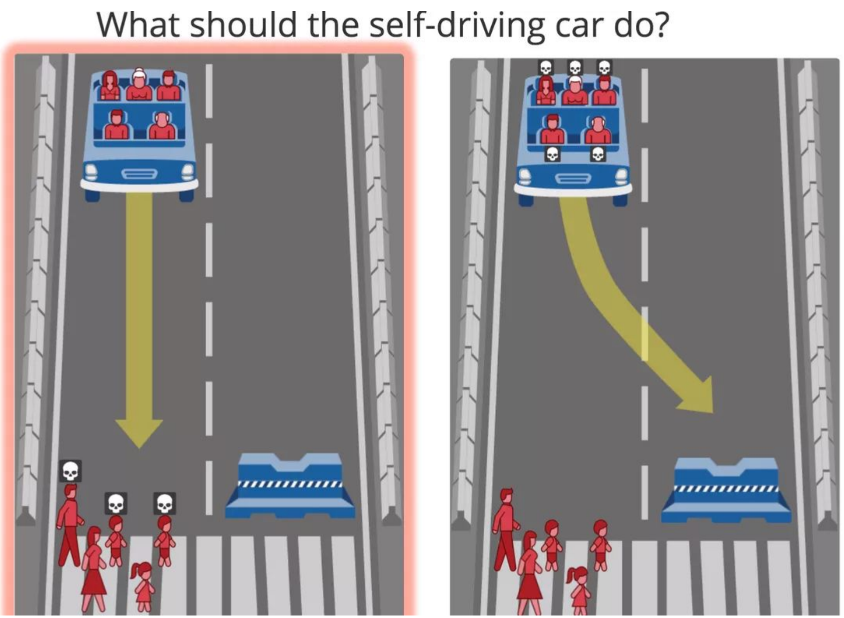
“the data subject should have the right [...] to obtain an explanation of the decision reached” - GDPR, Recital 71

Why is understanding the black box important ?

- GDPR

- Ethics

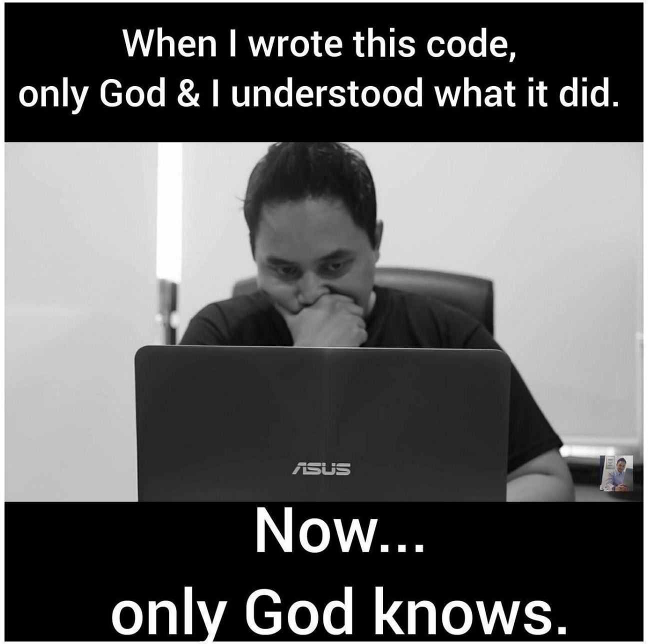
Awad, Edmond & Dsouza, Sohan & Kim, Richard & Schulz, Jonathan & Henrich, Joseph & Shariff, Azim & Bonnefon, Jean-François & Rahwan, Iyad. (2018). The Moral Machine Experiment. Nature. http://moralmachine.mit.edu/



What should the self-driving car do?

Why is understanding the black box important ?

- GDPR
- Ethics
- Users confidence



Why is understanding the black box important ?

- GDPR
- Ethics
- Users confidence
- Colleagues confidence

Why is understanding the black box important ?

- GDPR
- Ethics
- Users confidence
- Colleagues confidence

- Self confidence

Why is understanding the black box important ?

- GDPR
- Ethics
- Users confidence
- Colleagues confidence
- Self confidence

CONFIDENCE

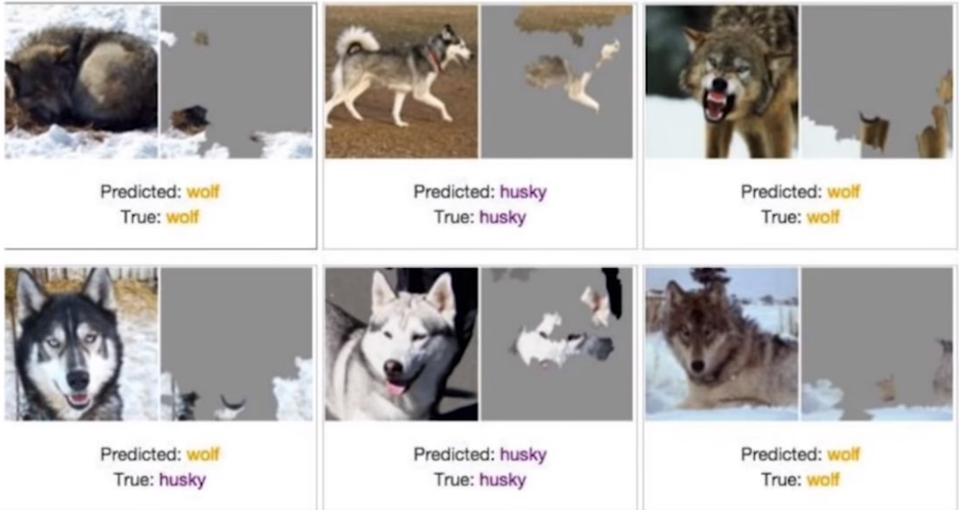
Why is understanding the black box important ?

- GDPR
- Ethics
- Users confidence
- Colleagues confidence
- Self confidence

CONFIDENCE

By cross validation !!!

Why is understanding the black box important ?



Predicted: wolf
True: wolf

Predicted: husky
True: husky

Predicted: wolf
True: wolf

Predicted: wolf
True: husky

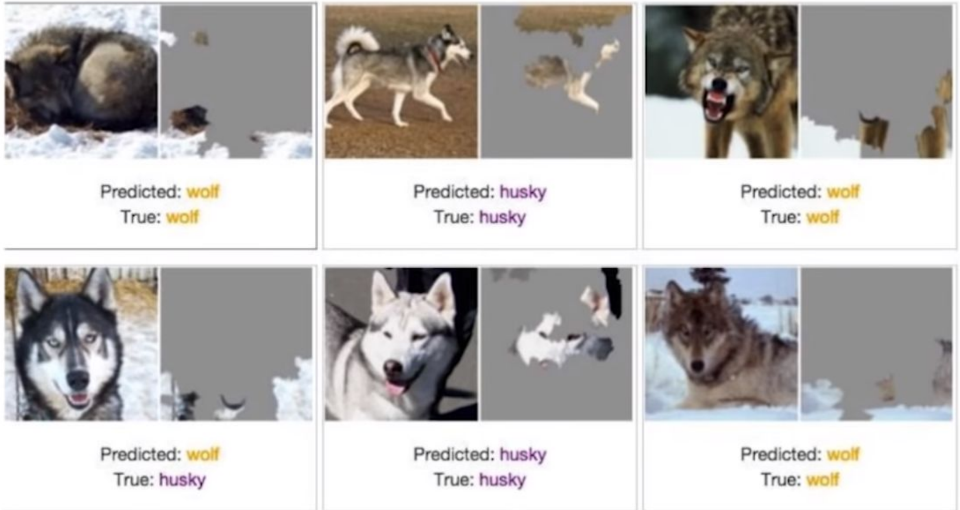
Predicted: husky
True: husky

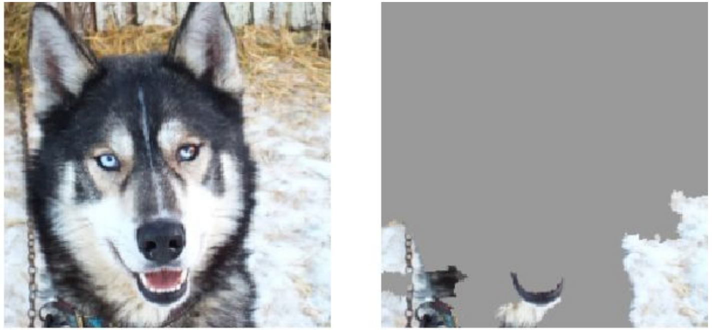
Predicted: wolf
True: wolf

Tulio Ribeiro, Marco & Singh, Sameer & Guestrin, Carlos. (2016). “Why Should I Trust You?”: Explaining the Predictions of Any Classifier

Why is understanding the black box important ?

Tulio Ribeiro, Marco & Singh, Sameer & Guestrin, Carlos. (2016). "Why Should I Trust You?": Explaining the Predictions of Any Classifier





(a) Husky classified as wolf

(b) Explanation

Why is understanding the black box important ?

- GDPR
- Ethics
- Users confidence
- Colleagues confidence
- Self confidence

- **Improve algoritms**

Current approaches

Current approaches

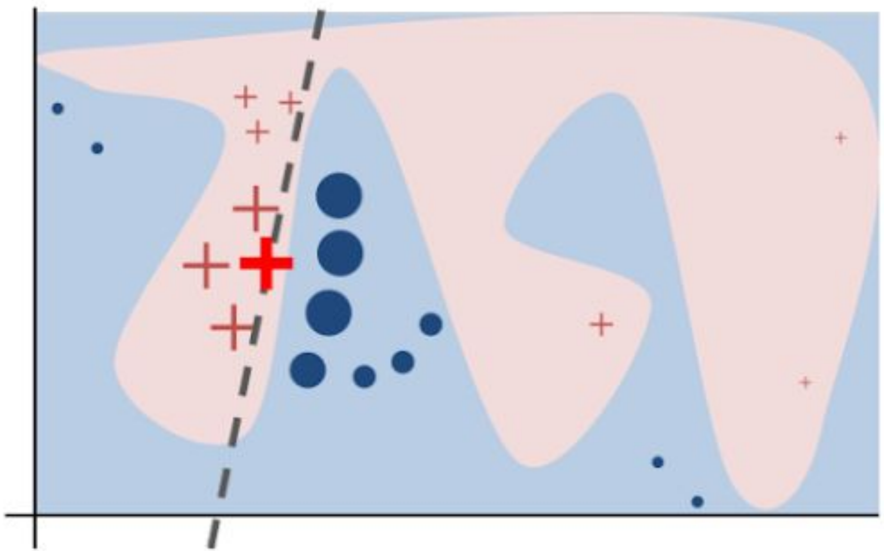
- *Local interpretability*
- *Global interpretability*

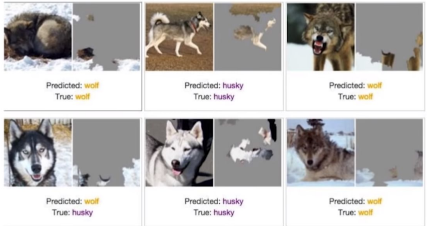
Current approaches

- ***Local interpretability:***

Explain one specific result

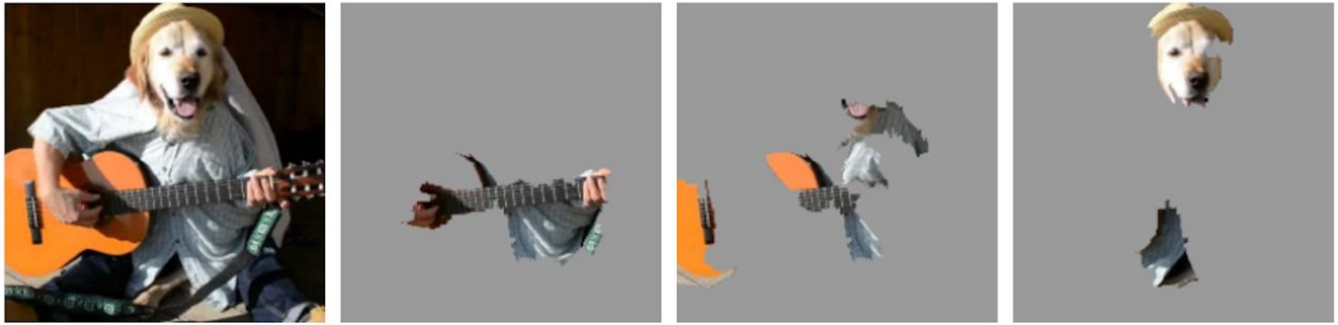
➢ *LIME - Local Interpretable Model-Agnostic Explanations*





Current approaches - Local interpretability

➢ *LIME - Local Interpretable Model-Agnostic Explanations*

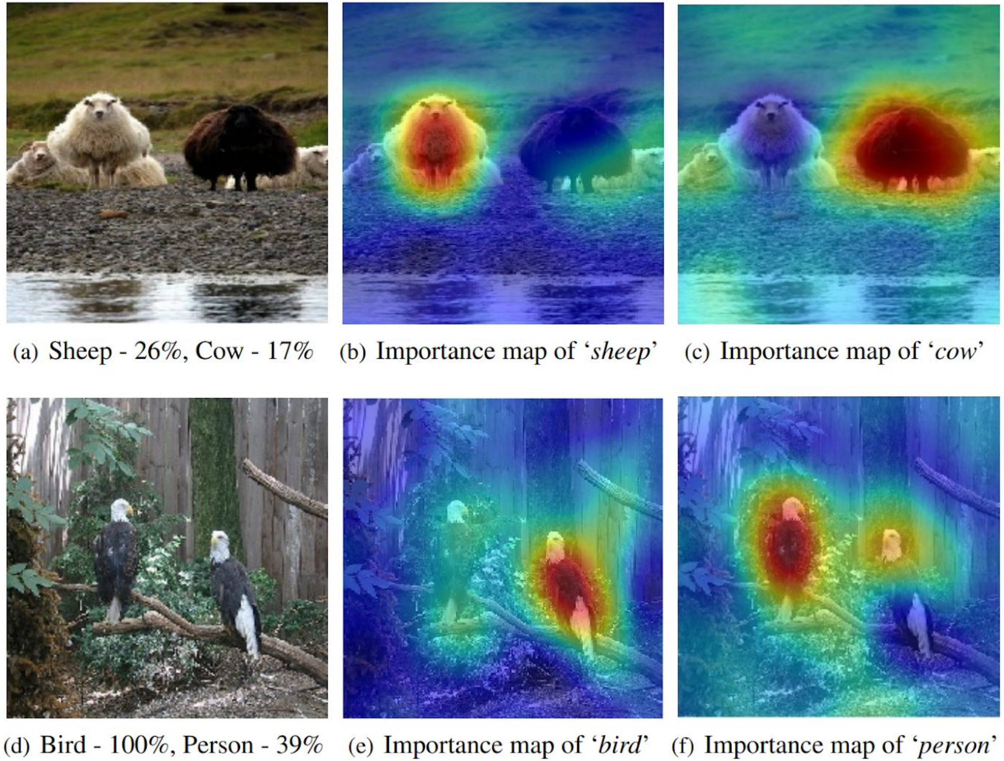


(a) Original Image (b) Explaining *Electric guitar* (c) Explaining *Acoustic guitar* (d) Explaining *Labrador*

Figure 4: Explaining an image classification prediction made by Google's Inception neural network. The top 3 classes predicted are "Electric Guitar" ($p = 0.32$), **"Acoustic guitar"** ($p = 0.24$) **and "Labrador"** ($p = 0.21$)

Current approaches - Local interpretability

➢ *RISE: Randomized Input Sampling for Explanation of Black-box Models*[1]



(a) Sheep - 26%, Cow - 17% (b) Importance map of '*sheep*' (c) Importance map of '*cow*'

(d) Bird - 100%, Person - 39% (e) Importance map of '*bird*' (f) Importance map of '*person*'

[1] Petsiuk, Vitali & Das, Abir & Saenko, Kate. (2018). RISE: Randomized Input Sampling for Explanation of Black-box Models. Boston University

Current approaches

- ***Global interpretability:***

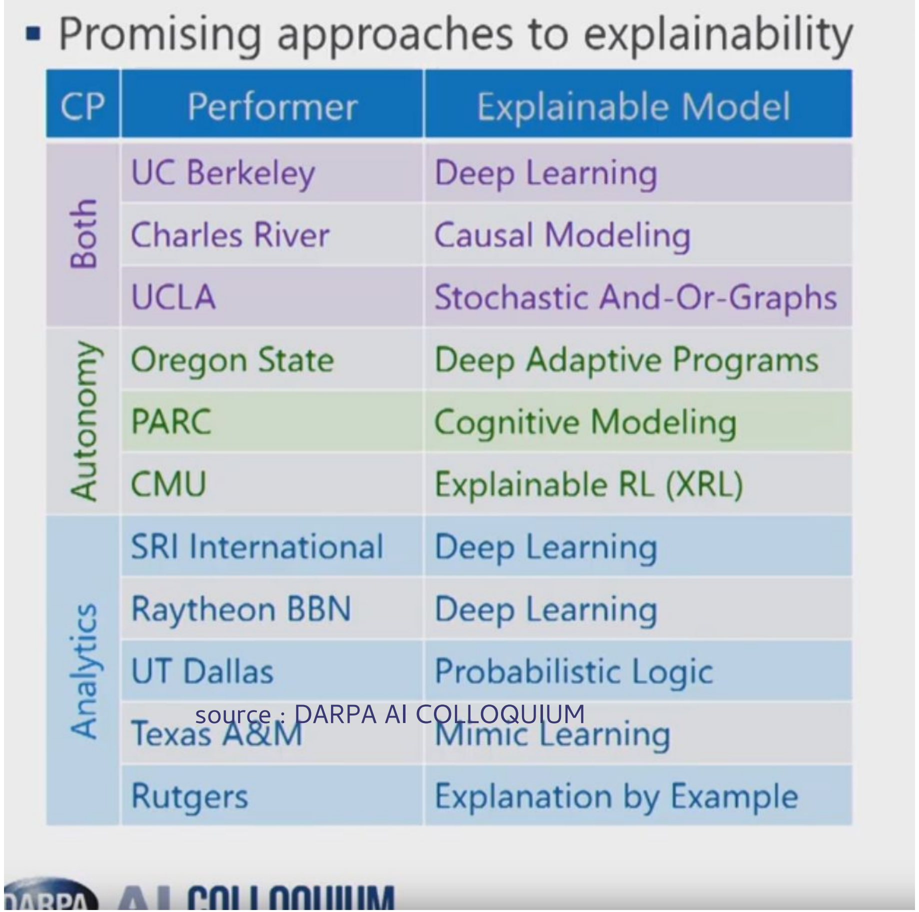
  Explain the whole model

  - ➢ *LIME suggest a kind of integration of the local interpretability*
  - ➢ *SHAP - A global model approximation based on linear combination of local approximations from different models*[1]

[1] Lundberg, Scott & Lee, Su-In. (2017). A Unified Approach to Interpreting Model Predictions - SHAP - University of Washington

Current approaches

- ***Hybrid approach***
- ***Others methods***
  - ➢ *Autoencoder*



- Promising approaches to explainability

CP	Performer	Explainable Model
Both	UC Berkeley	Deep Learning
	Charles River	Causal Modeling
	UCLA	Stochastic And-Or-Graphs
Autonomy	Oregon State	Deep Adaptive Programs
	PARC	Cognitive Modeling
	CMU	Explainable RL (XRL)
Analytics	SRI International	Deep Learning
	Raytheon BBN	Deep Learning
	UT Dallas	Probabilistic Logic
	Texas A&M	Mimic Learning
	Rutgers	Explanation by Example

source : DARPA AI COLLOQUIUM

Bibliography

ARTICLES :

- Tulio Ribeiro, Marco & Singh, Sameer & Guestrin, Carlos. (2016). **“Why Should I Trust You?”: Explaining the Predictions of Any Classifier**
- Petsiuk, Vitali & Das, Abir & Saenko, Kate. (2018). **RISE: Randomized Input Sampling for Explanation of Black-box Models**
- **GDPR - Recital 71**
- **For a meaningful artificial intelligence towards a french and european strategy**,Villani report, 2018
- Awad, Edmond & Dsouza, Sohan & Kim, Richard & Schulz, Jonathan & Henrich, Joseph & Shariff, Azim & Bonnefon, Jean-François & Rahwan, Iyad. (2018). **The Moral Machine Experiment**. Nature
- Lundberg, Scott & Lee, Su-In. (2017). **A Unified Approach to Interpreting Model Predictions**
- Zhou, Bolei & Sun, Yiyou & Bau, David & Torralba, Antonio. (2018). **Interpretable Basis Decomposition for Visual Explanation**

Other resources :

- **DARPA** - XAI - Literature Review
- **Fairness, Accountability, and Transparency in Machine Learning** (FAT) - https://www.fatml.org/

Clever Cloud Paris
137 rue vieille du temple 75003 Paris

Clever Cloud Nantes
3 rue de l'allier 44000 Nantes
02 85 52 07 69

https://www.clever-cloud.com

CONTACT

victor.ballu@clever-cloud.com