TELL ME WHY

BUILDING SAFE AND ROBUST MODELS: INTRODUCTION TO EXPLAINABILITY

Albert Calvo
@albertcalv
albertc@cs.upc.edu

UNIVERSITAT POLITÈCNICA DE CATALUNYA

WHOAMI

Bachelor in Informatics Engineering (UPC, 2016)

MSc in Innovation and Research (UPC & EPFL, 2018)

PhD in Computing (UPC)

PADS-UPC Research Group (Process Mining Data Science Group) https://www.cs.upc.edu/~pads-upc/

WHOAMI

Machine Learning in industry (Fraud Detection for Utilities)

Bundle of Datasets

Generalization about the data

Scoring

PROBLEM STATEMENT

User ID	Score
User 1	0.9555
User 2	0.9554
User …	…
User N	0.001

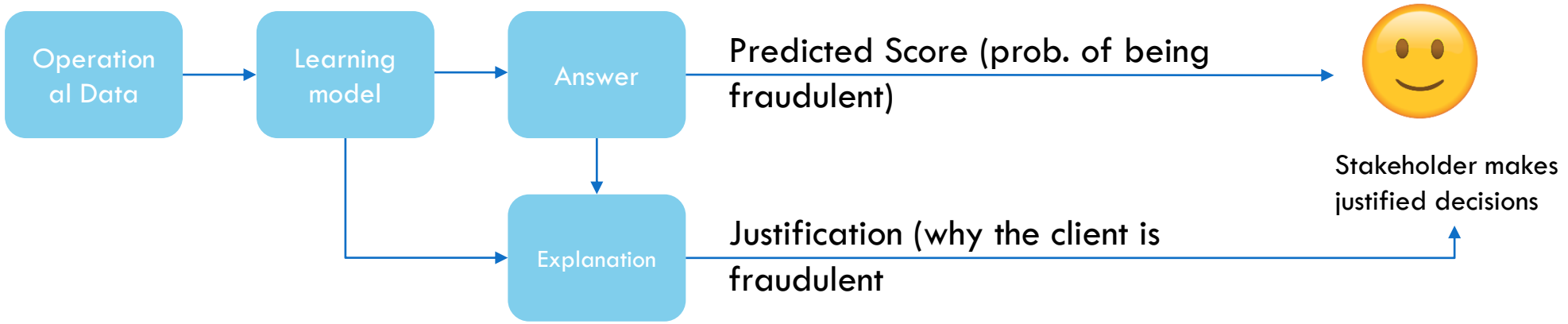
Stakeholder makes decision

Why did you that ?
When you succeed ?
When do you fail ?
When can I trust you ?
How do I correct an error ?

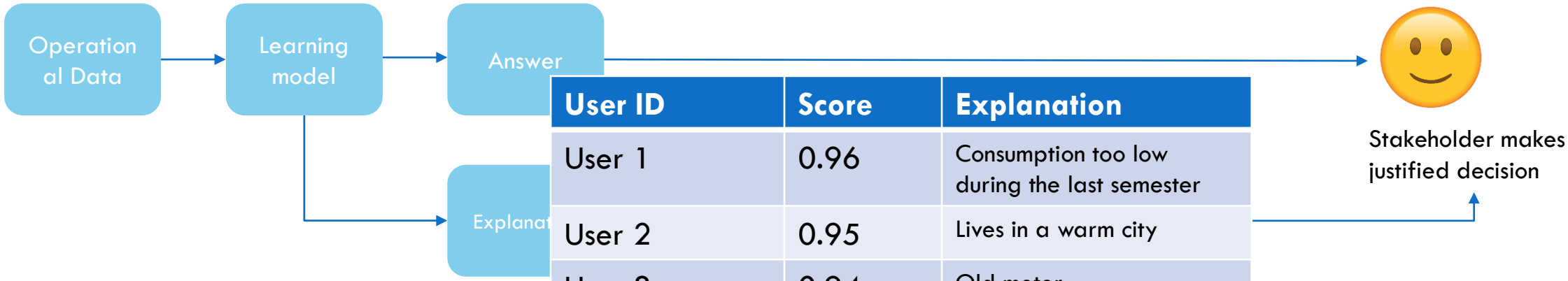
*Adapted from DARPA Slides

Problems of confidence and trust arise in industrial or sensitive models (high economic impact or lives)

PROBLEM STATEMENT



PROBLEM STATEMENT



I understand why
I understand why not
I know when you fail
I know when to trust you
I know why you erred

*Adapted from DARPA Slides

DEFINITION OF EXPLAINABILITY

· **Definition 1,** Science of comprehending what a model did, or might have done

[Leilani H. et altrum 2019]

· **Definition 2,** Ability to explain or to present understandable terms to a human

[Finale Doshi-Velez and Been Kim 2017]

· **Definition 3,** Use of machine-learning models for the extraction of relevant knowledge about domain relationships contained in data

[W. James Murdoch et altrum 2019]

EXPLAINABILITY GOALS

Explainability to achieve other important desiderata of ML systems

Fairness : protected groups are not somehow discriminated against

Privacy: means the method protects sensitive information in the data

Reliability & Robustness: ascertain whether algorithms reach certain levels of performance in the face of parameter of input variation

...

From : Towards A Rigorous Science of Interpretable Machine Learning Finale Doshi-Velez* and Been Kim*

AN ACTUAL DEMAND

Right to Explanation

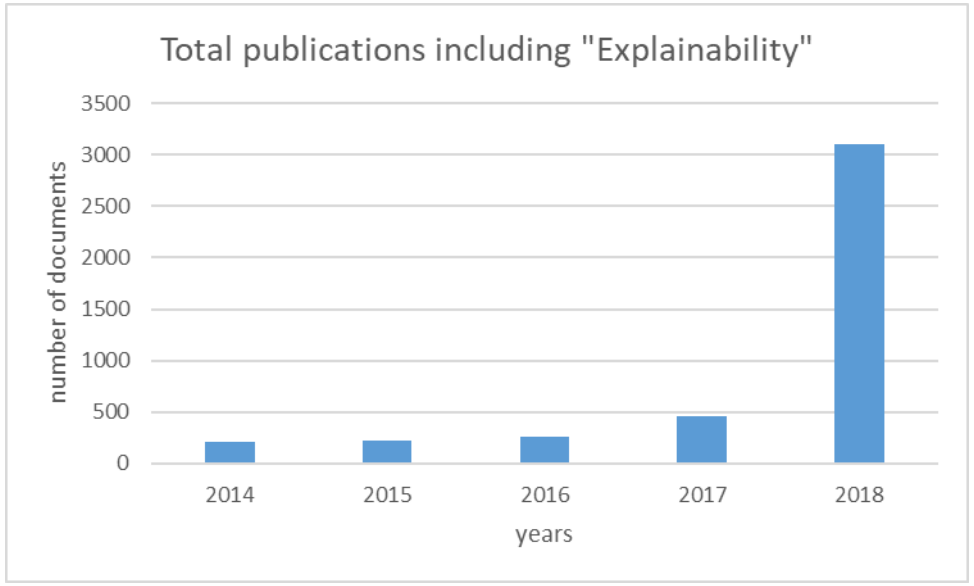
Human agency and oversight: AI systems should empower human beings, allowing them to make ***informed decisions*** *and fostering their fundamental rights.*

…

Transparency: the data, system and AI business models should be ***transparent****. Moreover, AI systems and their decisions should be* ***explained in a manner adapted to the stakeholder concerned****.*

https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai

AN ACTUAL DEMAND



Google Scholar

Explainability is included in top tier congress

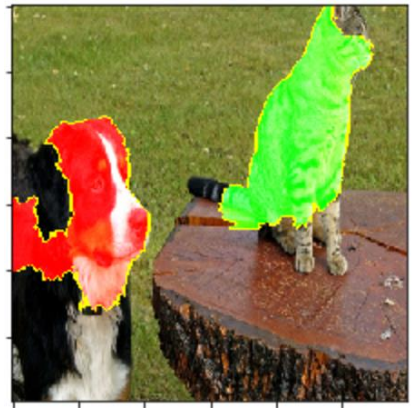
- ICML (International conference on Machine Learning), B++

- KDD (ACM International Conference On Knowledge Discovery and Data Mining), A++

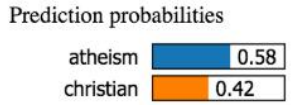
- NIPS (Neural Information Processing Systems), A++

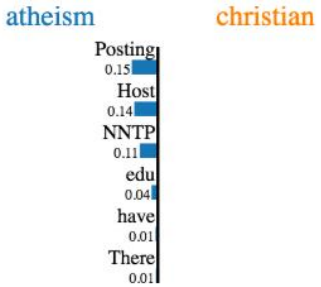
- ECAI (European Conference on Artificial Intelligence), A

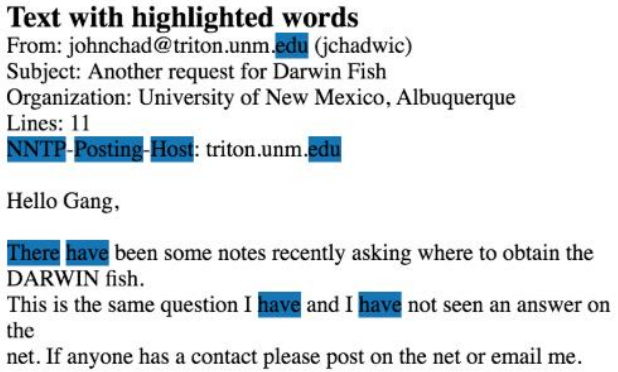
EXAMPLES OF EXPLAINABILITY

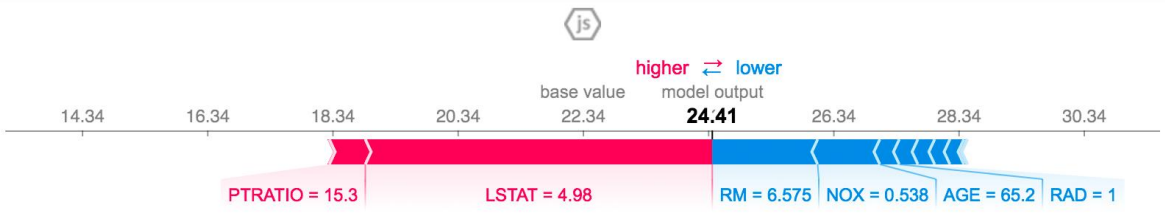


Explanations in Images









Explanations in Text

User ID	Score	Explanation
User 1	0.96	Consumption too low during the last semester
User 2	0.95	Lives in a warm city
User 3	0.94	Old meter
User 4	0.4	New client

Explanations in binary classification

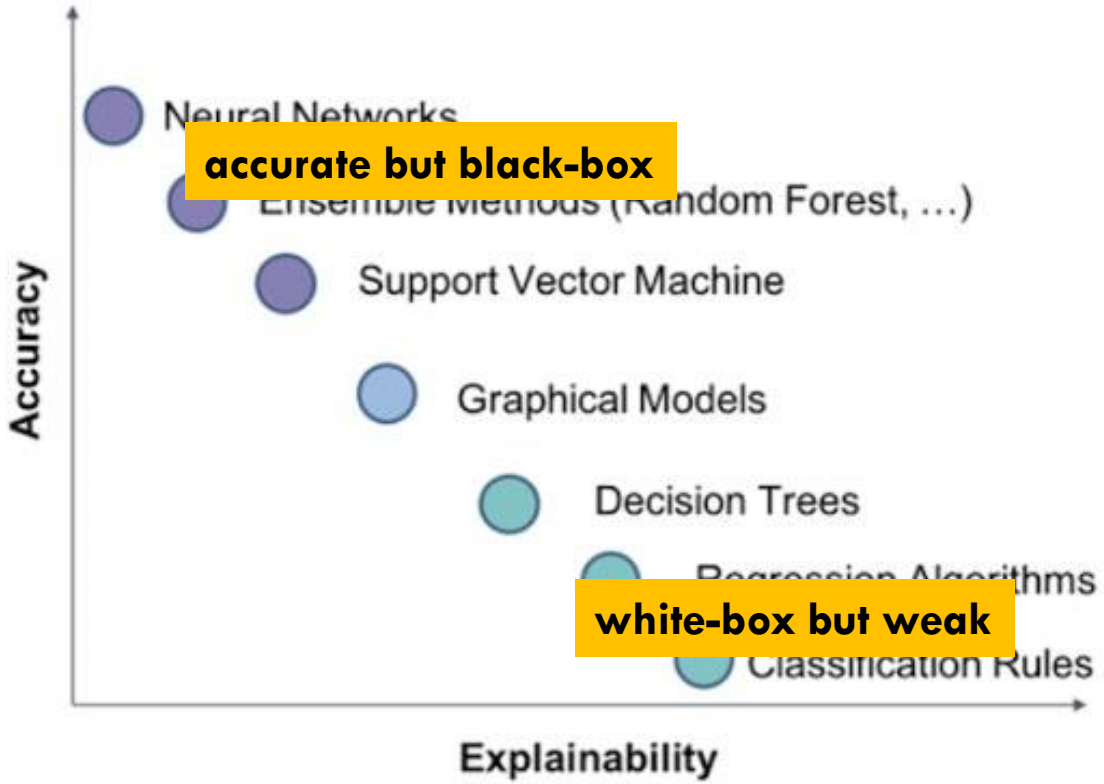
TOOLS FOR EXPLAINABILITY

Two different approaches:

- Approach 1 : Post-hoc explanations
  - Individual prediction explanations: perturbations of single points or contributions
  - Global prediction explanations: summary plots, dependence plots etc.
- Approach 2 : Build Interpretable models
  - Decisions Trees, Decisions Rules etc.

TOOLS FOR EXPLAINABILITY

What model to choose ?



Accuracy – Explainability trade-off

Chapter II
Black Box Techniques
LIME and SHAP as Post-hoc analysis

Chapter I
White Box Techniques
Classification Rules and Decision Trees

THANKS FOR YOUR ATTENTION

GitHub Repository:
https://github.com/albertcalv/tellmewhy

Albert Calvo
@albertcalv
albertc@cs.upc.edu