TELL ME WHY

BUILDING SAFE AND ROBUST MODELS: INTRODUCTION TO EXPLAINABILITY

Albert Calvo

@albertcalv

@ albertc@cs.upc.edu

UNIVERSITAT POLITÈCNICA
DE CATALUNYA
UPC

WHOAMI

Bachelor in Informatics Engineering (UPC, 2016)

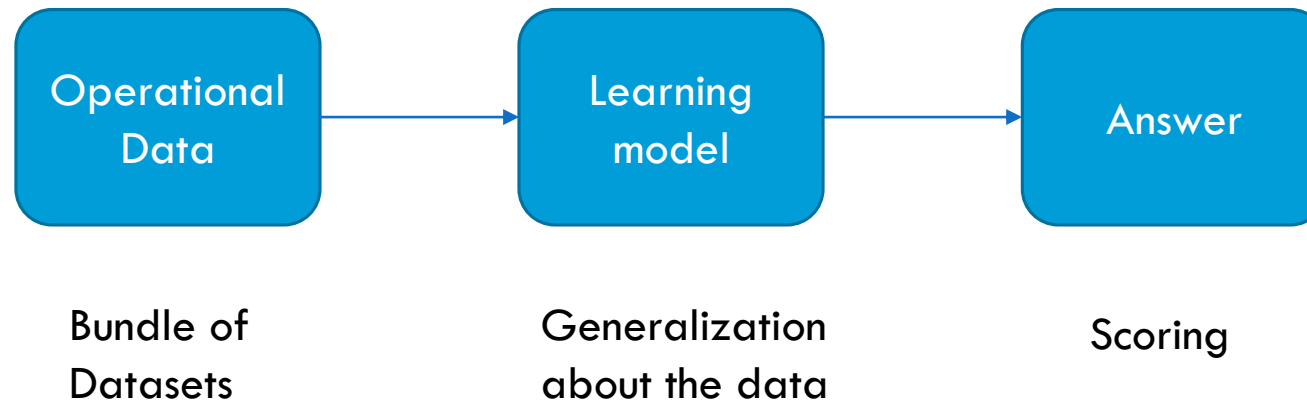MSc in Innovation and Research (UPC & EPFL, 2018)

PhD in Computing (UPC)

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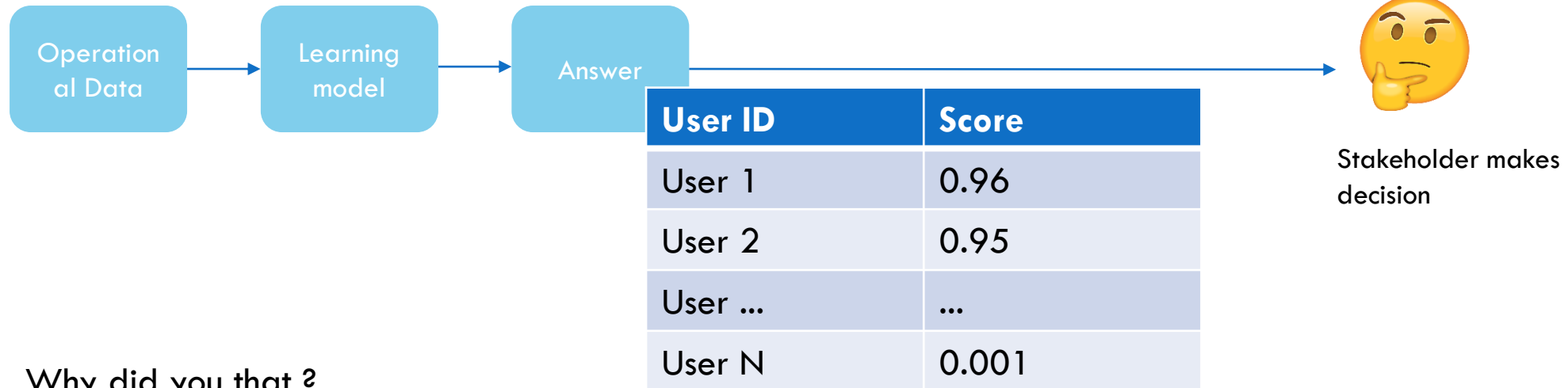
Science Group) https://www.cs.upc.edu/~pads-upc/

WHOAMI

Machine Learning in industry (Fraud Detection for Utilities)

Operational Data	→	Learning model	→	Answer
Bundle of Datasets		Generalization about the data		Scoring

PROBLEM STATEMENT

Operation al Data	→	Learning model	→	Answer	→

🤔

Stakeholder makes decision

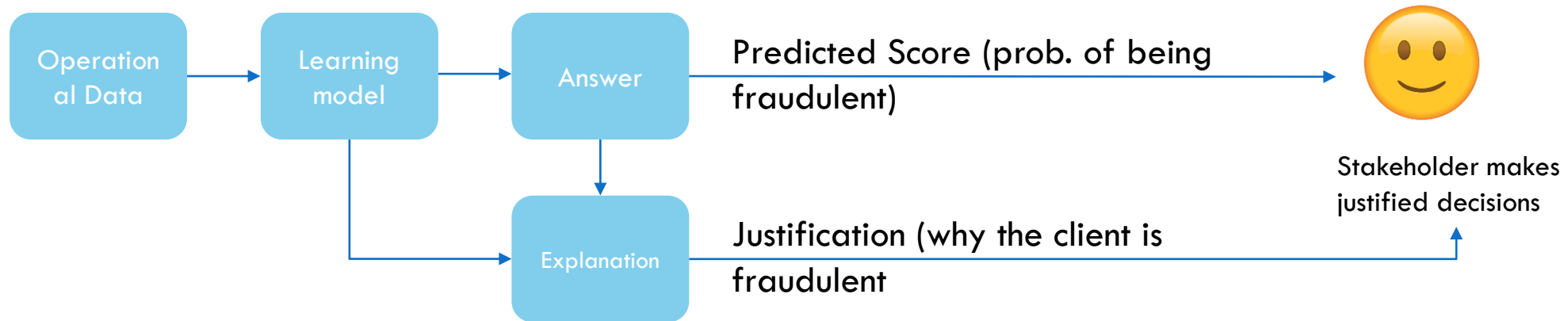| User ID	Score
User 1	0.96
User 2	0.95
User ...	...
User N	0.001

Why did you that ?
When you succeed ?
When do you fail ?
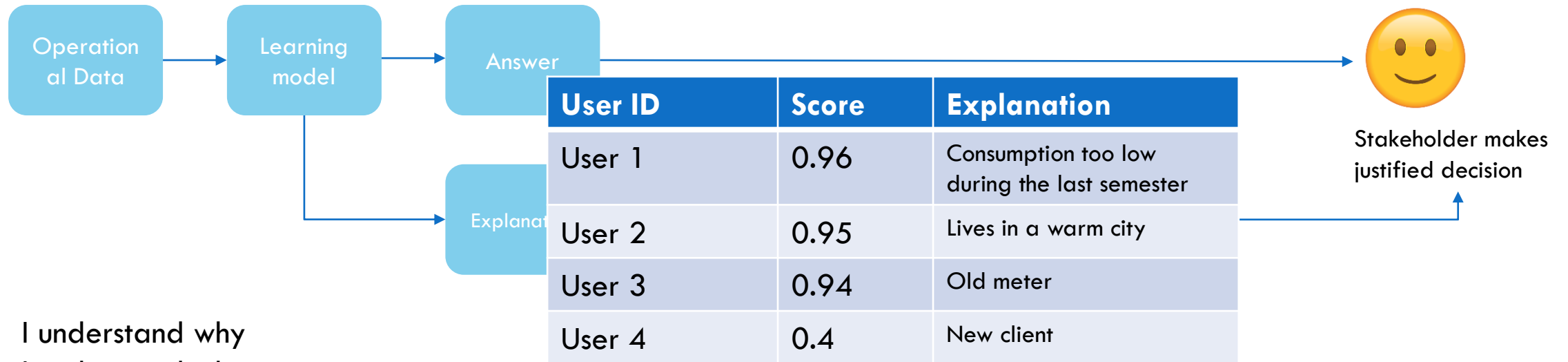When can I trust you ?
How do I correct an error ?
*Adapted from DARPA Slides

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

PROBLEM STATEMENT

Operational Data → Learning model → Answer

Predicted Score (prob. of being fraudulent)

Explanation

Justification (why the client is fraudulent

Stakeholder makes justified decisions

PROBLEM STATEMENT

| Operation al Data | → | Learning model | → | Answer | → |

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

Explanat...

Stakeholder makes justified decision

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*
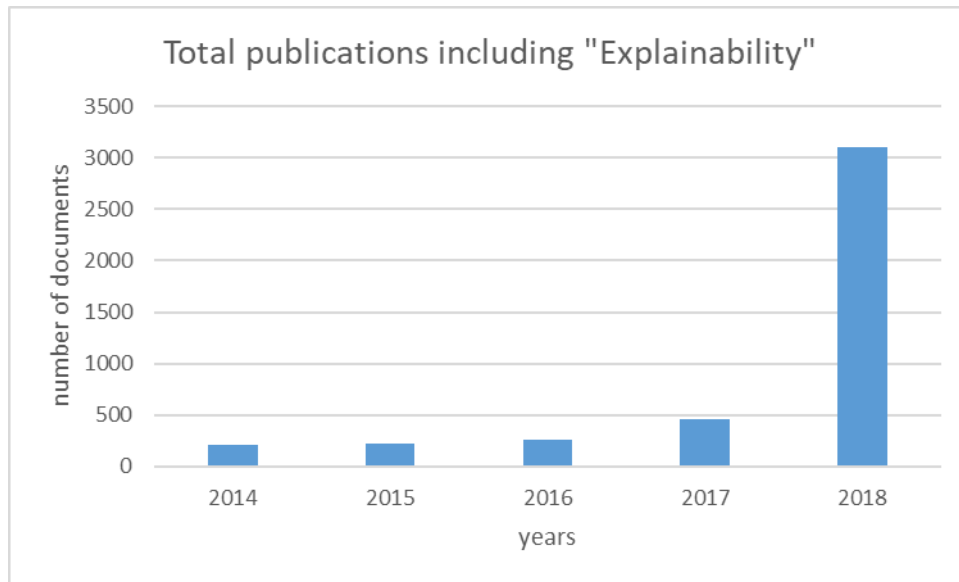
AN ACTUAL DEMAND

*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**.*

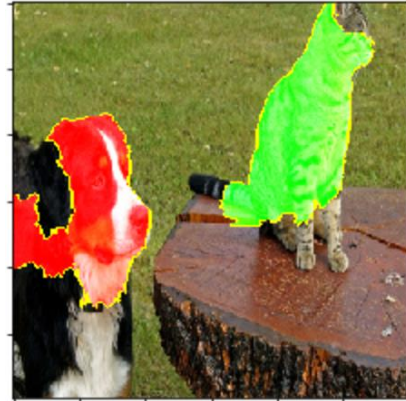https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai

AN ACTUAL DEMAND

Total publications including "Explainability"

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
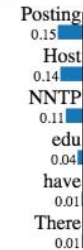
EXAMPLES OF EXPLAINABILITY

Explanations in Images

Prediction probabilities

| atheism | 0.58 |
| christian | 0.42 |

atheism   christian

Posting 0.15
Host 0.14
NNTP 0.11
edu 0.04
have 0.01
There 0.01

Text with highlighted words

From: johnchad@triton.unm.edu (jchadwic)
Subject: Another request for Darwin Fish
Organization: University of New Mexico, Albuquerque
Lines: 11
NNTP-Posting-Host: triton.unm.edu

Hello Gang,

There have been some notes recently asking where to obtain the DARWIN fish.
This is the same question I have and I have not seen an answer on the
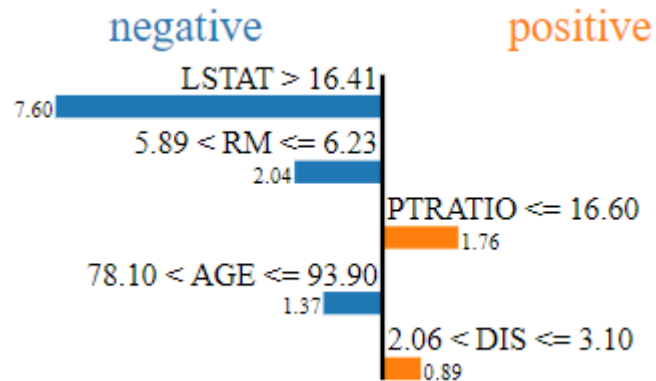net. If anyone has a contact please post on the net or email me.

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

higher ⇄ lower
base value   model output
14.34   16.34   18.34   20.34   22.34   **24.41**   26.34   28.34   30.34

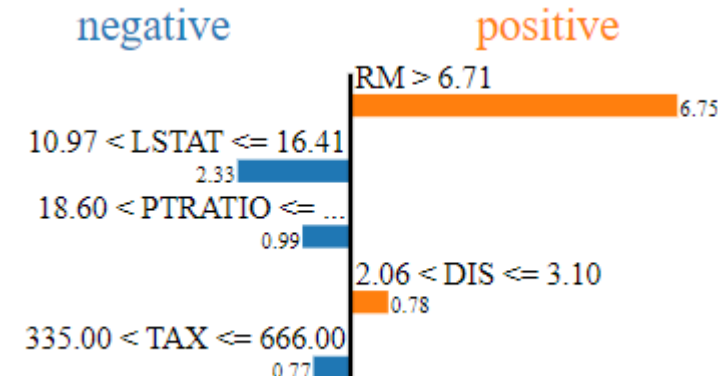PTRATIO = 15.3   LSTAT = 4.98   RM = 6.575 NOX = 0.538 AGE = 65.2 RAD = 1

Explanations in binary classification

OPEN CHALLENGES (TOO MANY!!)

Explanation 1, predicting instance A using model ∂

Explanation 2, predicting instance A using model Ω

How to measure explainability?

Is better Explanation 1 or Explanation 2

How to represent this explanations?

Plot, Summary, Interactions between features ...

How integrate humans in the ML pipeline ?

How to Involve the human during the learning stage

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 ?
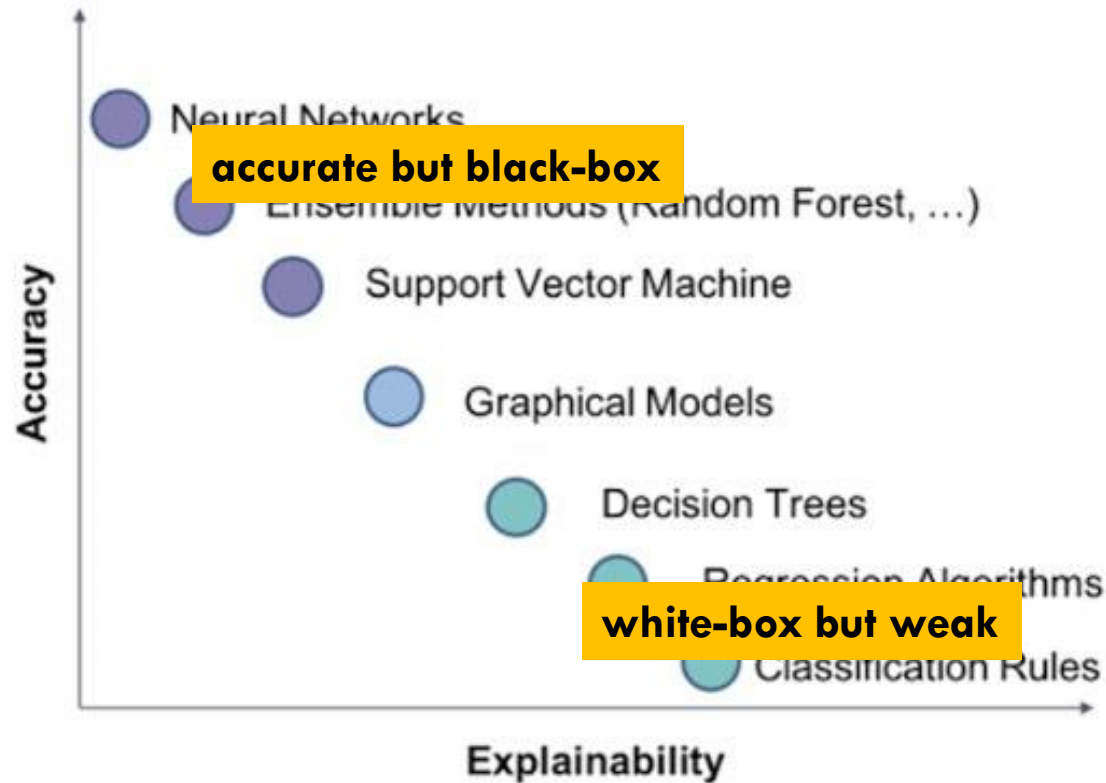
Accuracy – Explainability trade-off

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

Chapter I
White Box Techniques
Classification Rules and Decision Trees

GitHub Repository:

https://github.com/albertcalv/tellmewhy

WHITE BOX MODELS

HANDS-ON I

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

BLACK BOX MODELS | HANDS-ON II - III

THANKS FOR YOUR ATTENTION

Albert Calvo

@albertcalv

✉ albertc@cs.upc.edu