# Big Data Analytics

Fosca Giannotti and Luca Pappalardo

http://didawiki.di.unipi.it/doku.php/bigdataanalytics/bda/

**DIPARTIMENTO DI INFORMATICA - Università di Pisa**
**anno accademico 2019/2020**

# Explainable AI: From Theory to Motivation, Applications and Challenges

# What is "Explainable AI" ?

- Explainable-AI explores and investigates methods to produce or complement AI models to make accessible and interpretable the internal logic and the outcome of the algorithms, making such process understandable by humans.

# What is ”Explainable AI” ?

Explicability, understood as incorporating both intelligibility (“how does it work?” for non-experts, e.g., patients or business customers, and for experts, e.g., product designers or engineers) and accountability (“who is responsible for”).

- 5 core principles for ethical AI:
  - beneficence, non-maleficence, autonomy, and justice
  - a new principle is needed in addition: explicability

# Tutorial Outline (1)

- **Motivating Examples**
- **Explanation in AI**
  - Explanations in different AI fields
  - The Role of Humans
  - Evaluation Protocols & Metrics
- **Explainable Machine Learning**
  - What is a Black Box?
  - Interpretable, Explainable, and Comprehensible Models
  - Open the Black Box Problems
- **Guidelines** for explaining AI systems

# Properties of Interpretable ML Models

# Interpretability

- To ***interpret*** means to give or provide the meaning or to explain and present in understandable terms some concepts.

- In data mining and machine learning, interpretability is the ***ability to explain*** or to provide the meaning ***in understandable terms to a human***.

- https://www.merriam-webster.com/
- Finale Doshi-Velez and Been Kim. 2017. ***Towards a rigorous science of interpretable machine learning***. arXiv:1702.08608v2.

# Dimensions of Interpretability

- ***Global and Local Interpretability***:
  - *Global*: understanding the whole logic of a model
  - *Local*: understanding only the reasons for a specific decision

- ***Time Limitation***: the time that the user can spend for understanding an explanation.

- ***Nature of User Expertise***: users of a predictive model may have different background knowledge and experience in the task. The nature of the user expertise is a key aspect for interpretability of a model.

# Desiderata of an Interpretable Model

- ***Interpretability*** *(or* comprehensibility*)*: to which extent the model and/or its predictions are human understandable. Is measured with the *complexity* of the model.

- ***Fidelity***: to which extent the model imitate a black-box predictor.

- ***Accuracy***: to which extent the model predicts unseen instances.

- Alex A. Freitas. 2014. ***Comprehensible classification models: A position paper***. ACM SIGKDD Explor. Newslett.

# Desiderata of an Interpretable Model

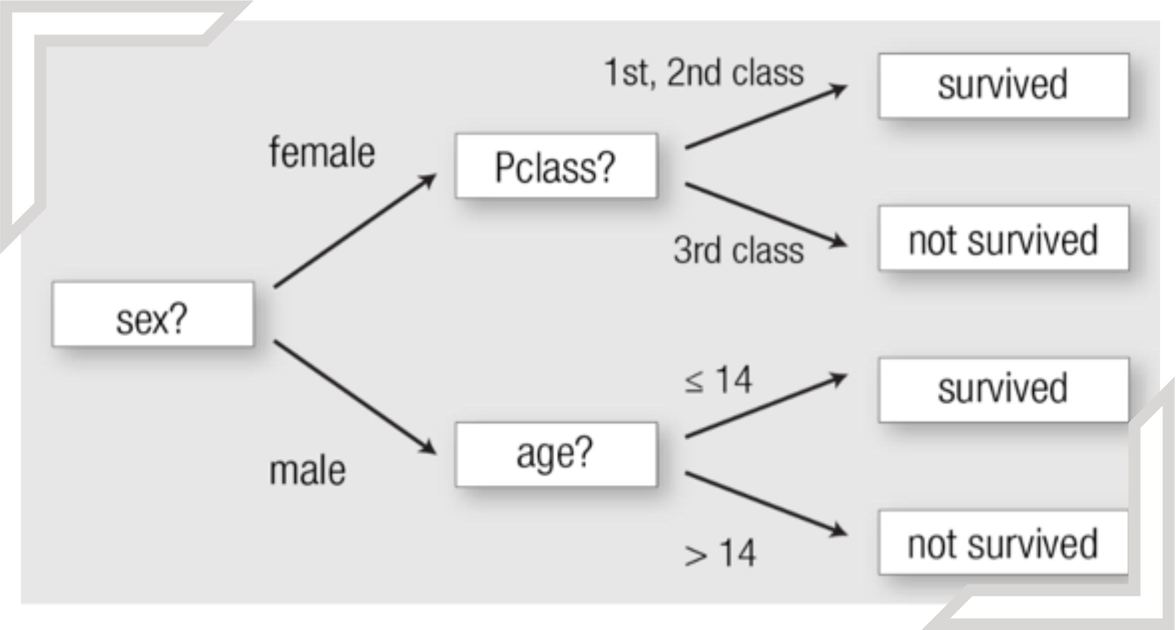
- ***Fairness***: the model guarantees the protection of groups against discrimination.
- ***Privacy***: the model does not reveal sensitive information about people.
- ***Respect Monotonicity***: the increase of the values of an attribute either increase or decrease in a monotonic way the probability of a record of being member of a class.
- ***Usability***: an interactive and queryable explanation is more usable than a textual and fixed explanation.


- Andrea Romei and Salvatore Ruggieri. 2014. ***A multidisciplinary survey on discrimination analysis***. Knowl. Eng.
- Yousra Abdul Alsahib S. Aldeen, Mazleena Salleh, and Mohammad Abdur Razzaque. 2015. ***A comprehensive review on privacy preserving data mining***. SpringerPlus .
- Alex A. Freitas. 2014. ***Comprehensible classification models: A position paper***. ACM SIGKDD Explor. Newslett.

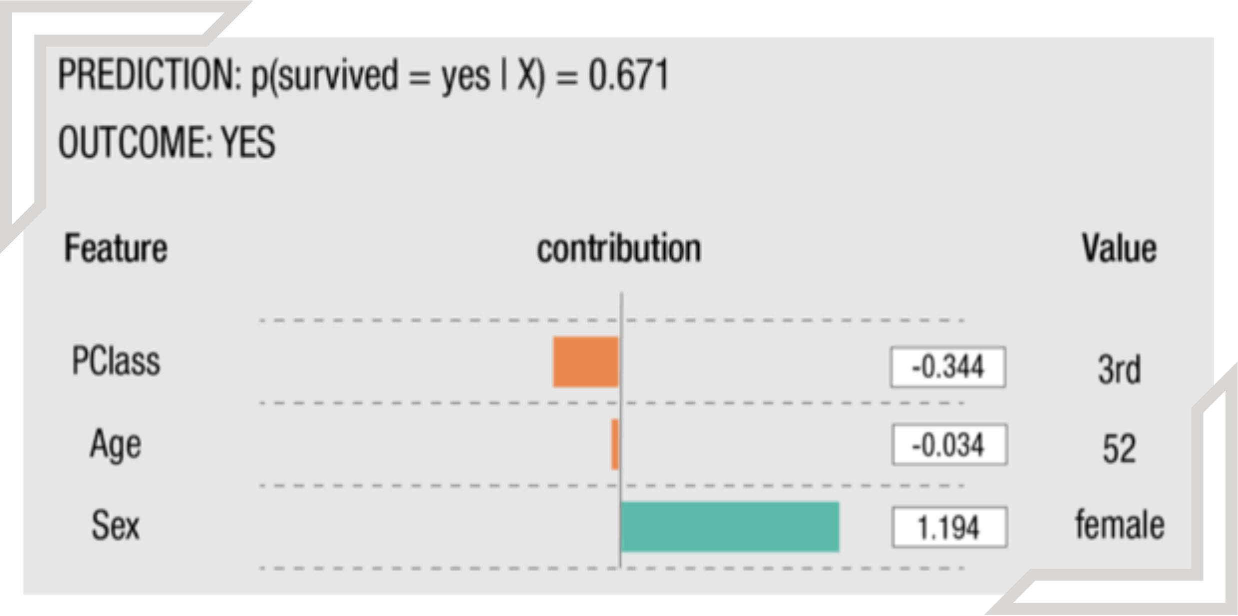
# Desiderata of an Interpretable Model

- ***Reliability and Robustness***: the interpretable model should maintain high levels of performance independently from small variations of the parameters or of the input data.
- ***Causality:*** controlled changes in the input due to a perturbation should affect the model behavior.
- ***Scalability:*** the interpretable model should be able to scale to large input data with large input spaces.
- ***Generality:*** the model should not require special training or restrictions.

# Recognized Interpretable Models




Decision Tree




Linear Model

$$\text{if } condition_1 \wedge condition_2 \wedge condition_3 \text{ then } outcome$$

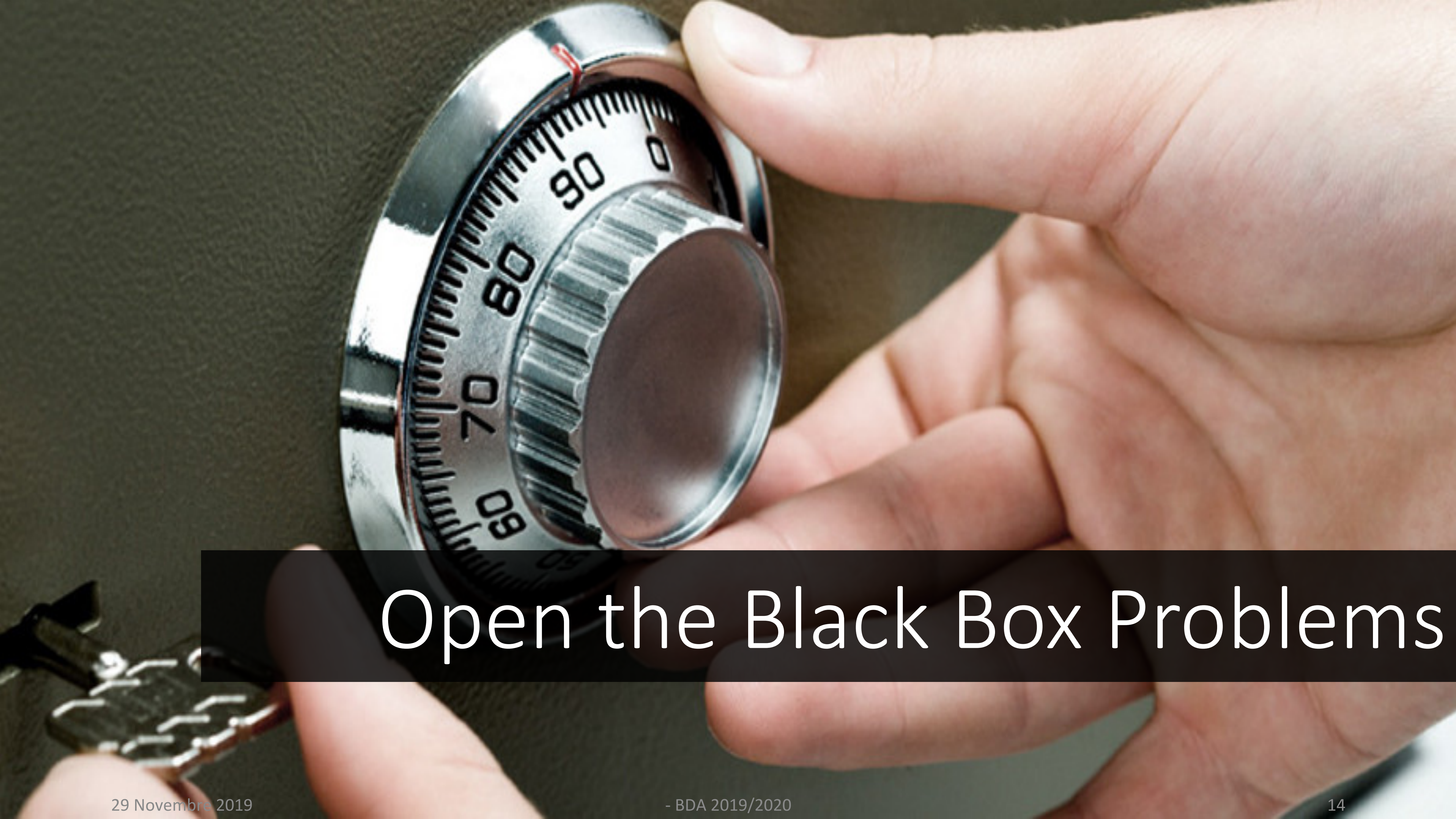
Rules

# Complexity

- Opposed to *interpretability*.
- Is only related to the model and not to the training data that is unknown.
- Generally estimated with a rough approximation related to the ***size*** of the interpretable model.

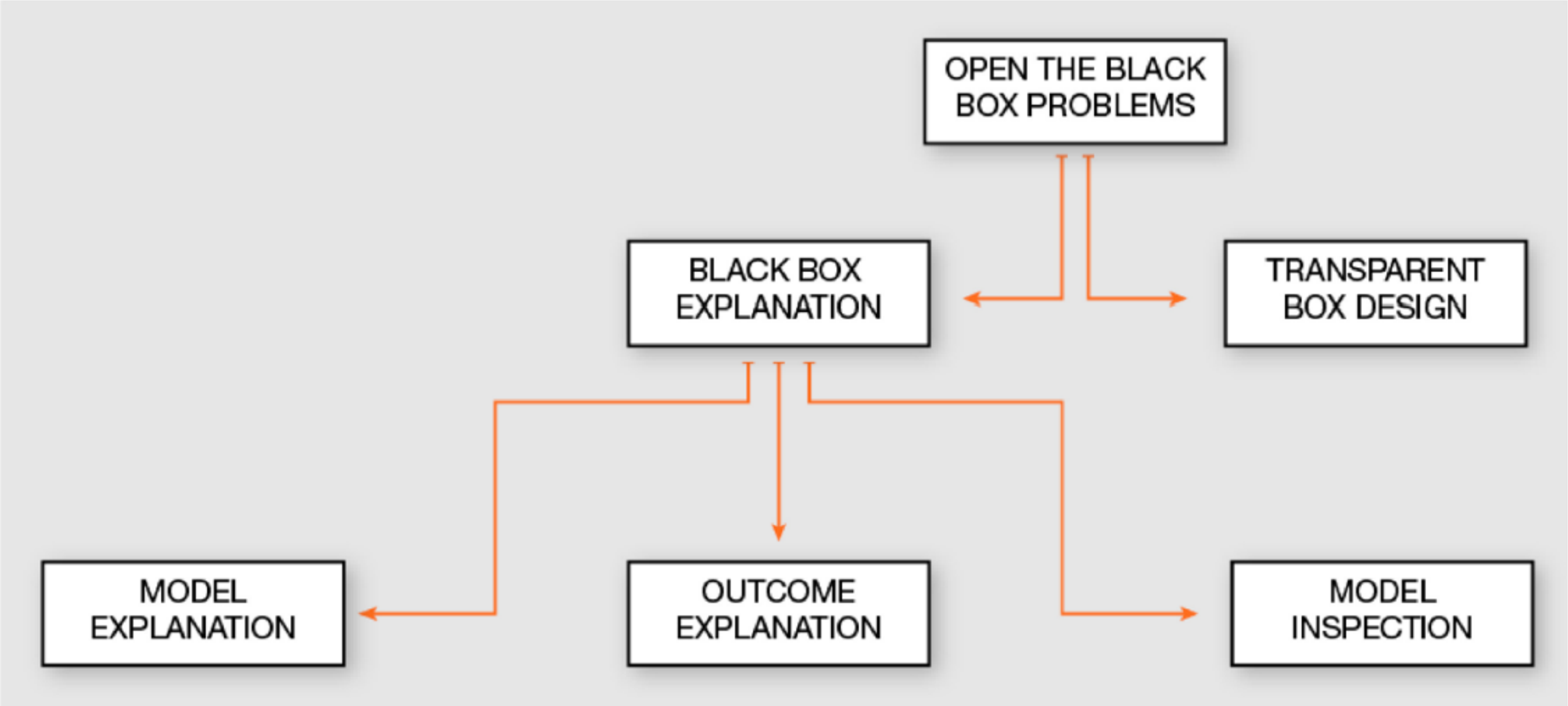
- Linear Model: number of non zero weights in the model.
- Rule: number of attribute-value pairs in condition.
- Decision Tree: estimating the complexity of a tree can be hard.


- Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. ***Why should i trust you?: Explaining the predictions of any classifier***. KDD.
- Houtao Deng. 2014. ***Interpreting tree ensembles with intrees***. arXiv preprint arXiv:1408.5456.
- Alex A. Freitas. 2014. ***Comprehensible classification models: A position paper***. ACM SIGKDD Explor. Newslett.




# Open the Black Box Problems

# Problems Taxonomy



# Black Boxes

- Neural Network (***NN***)
- Tree Ensemble (***TE***)
- Support Vector Machine (***SVM***)
- Deep Neural Network (***DNN***)

# Types of Data

Table of baby-name data
(baby-2010.csv)

| name | rank | gender | year |
|---|---|---|---|
| Jacob | 1 | boy | 2010 |
| Isabella | 1 | girl | 2010 |
| Ethan | 2 | boy | 2010 |
| Sophia | 2 | girl | 2010 |
| Michael | 3 | boy | 2010 |

Field names

One row (4 fields)

2000 rows all told

Tabular

**(TAB)**

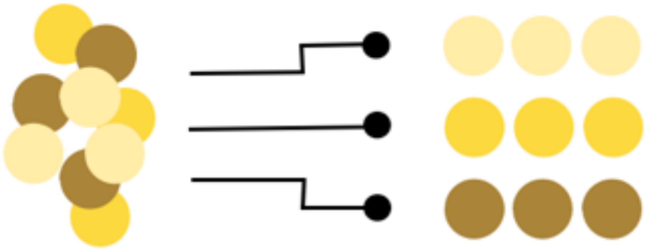
Images

**(IMG)**


Text

**(TXT)**

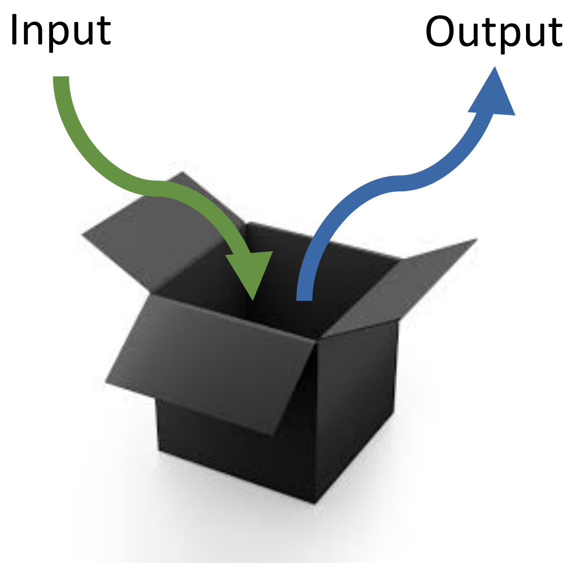
# Explanators



- Decision Tree (***DT***)
- Decision Rules (***DR***)
- Features Importance (***FI***)
- Saliency Mask (***SM***)
- Sensitivity Analysis (***SA***)
- Partial Dependence Plot (***PDP***)
- Prototype Selection (***PS***)
- Activation Maximization (***AM***)

# Reverse Engineering

- The name comes from the fact that we can only ***observe*** the ***input*** and ***output*** of the black box.
- Possible actions are:
  - ***choice*** of a particular comprehensible predictor
  - querying/auditing the black box with input records created in a controlled way using ***random perturbations*** w.r.t. a certain prior knowledge (e.g. train or test)
- It can be ***generalizable or not***:
  - Model-Agnostic
  - Model-Specific



# Solving The Model Explanation Problem

| Name | Ref. | Authors | Year | Explanator | Black Box | Data Type | General | Random | Examples | Code | Dataset |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Trepan | [22] | Craven et al. | 1996 | DT | NN | TAB | ✓ | | | | ✓ |
| – | [57] | Krishnan et al. | 1999 | DT | NN | TAB | ✓ | | ✓ | | ✓ |
| DecText | [12] | Boz | 2002 | DT | NN | TAB | ✓ | ✓ | | | ✓ |
| GPDT | [46] | Johansson et al. | 2009 | DT | NN | TAB | ✓ | ✓ | ✓ | | ✓ |
| Tree Metrics | [17] | Chipman et al. | 1998 | DT | TE | TAB | | | | | ✓ |
| CCM | [26] | Domingos et al. | 1998 | DT | TE | TAB | ✓ | ✓ | | | ✓ |
| – | [34] | Gibbons et al. | 2013 | DT | TE | TAB | ✓ | ✓ | | | |
| STA | [140] | Zhou et al. | 2016 | DT | TE | TAB | | ✓ | | | |
| CDT | [104] | Schetinin et al. | 2007 | DT | TE | TAB | | | ✓ | | |
| – | [38] | Hara et al. | 2016 | DT | TE | TAB | | ✓ | ✓ | | ✓ |
| TSP | [117] | Tan et al. | 2016 | DT | TE | TAB | | | | | ✓ |
| Conj Rules | [21] | Craven et al. | [illegible] | DR | NN | TAB | | | | | |
| G-REX | [44] | Johansson et al. | 2003 | DR | NN | TAB | ✓ | ✓ | ✓ | | |
| REFNE | [141] | Zhou et al. | 2003 | DR | NN | TAB | ✓ | ✓ | ✓ | | ✓ |
| RxREN | [6] | Augasta et al. | 2012 | DR | NN | TAB | | ✓ | ✓ | | ✓ |

# Global Model Explainers

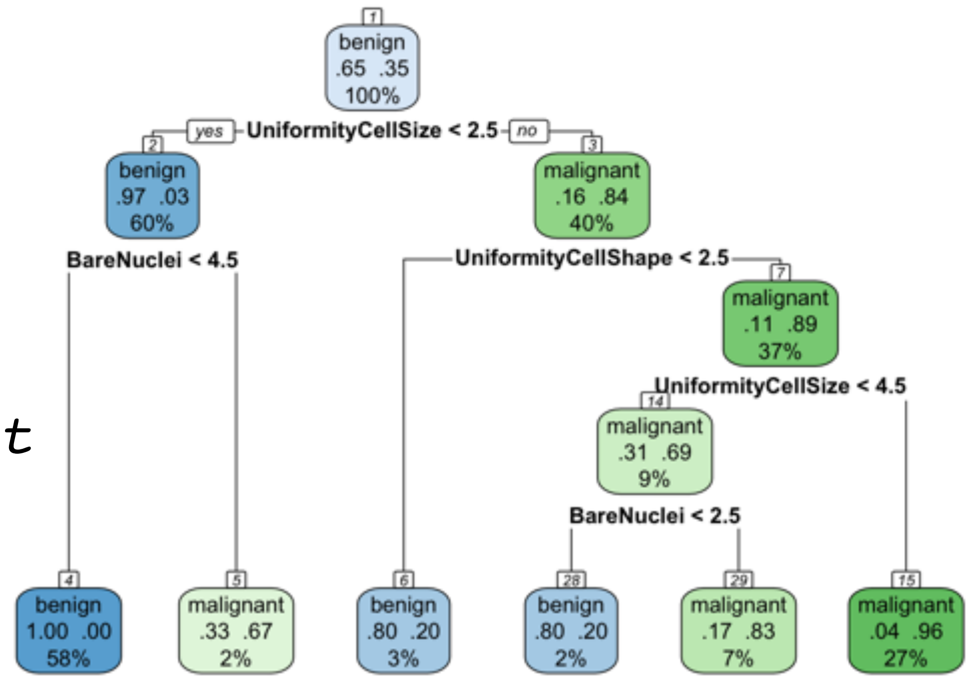
- Explanator: DT
  - Black Box: NN, TE
  - Data Type: TAB
- Explanator: DR
  - Black Box: NN, SVM, TE
  - Data Type: TAB
- Explanator: FI
  - Black Box: AGN
  - Data Type: TAB

$R_1$ : IF(Outlook = Sunny) AND (Windy= False) THEN Play=Yes
$R_2$ : IF(Outlook = Sunny) AND (Windy= True) THEN Play=No
$R_3$ : IF(Outlook = Overcast) THEN Play=Yes
$R_4$ : IF(Outlook = Rainy) AND (Humidity= High) THEN Play=No
$R_5$ : IF(Outlook = Rainy) AND (Humidity= Normal) THEN Play=Yes

# Trepan – DT, NN, TAB

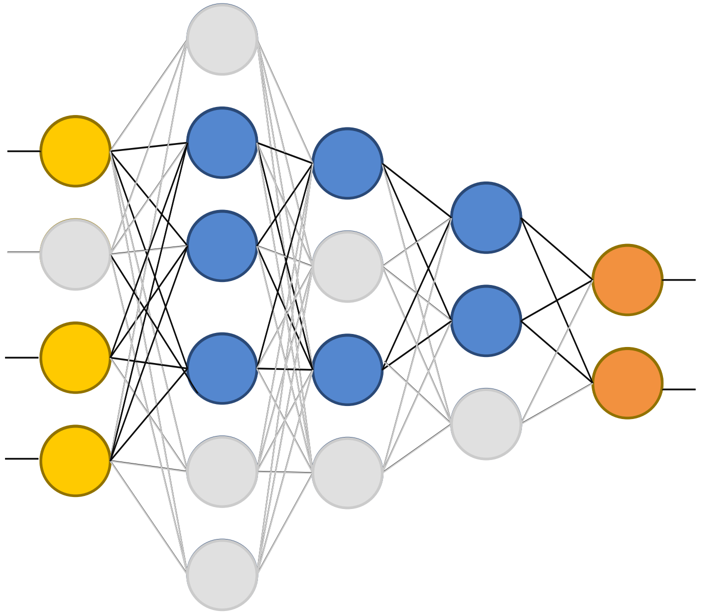
*black box auditing*

```
T = root_of_the_tree()
Q = <T, X, {}>
while Q not empty & size(T) < limit
      N, X_N, C_N  = pop(Q)
      Z_N  = random(X_N, C_N)
      y_Z  = b(Z), y = b(X_N)
      if same_class(y ∪ y_Z)
            continue
      S = best_split(X_N ∪ Z_N, y ∪ y_Z)
      S'= best_m-of-n_split(S)
      N = update_with_split(N, S')
      for each condition c in S'
            C = new_child_of(N)
            C_C  = C_N ∪ {c}
            X_C  = select_with_constraints(X_N, C_N)
            put(Q, <C, X_C, C_C>)
```




- Mark Craven and JudeW. Shavlik. 1996. ***Extracting tree-structured representations of trained networks***. NIPS.

# RxREN – DR, NN, TAB



```
prune insignificant neurons
for each significant neuron
  for each outcome
    compute mandatory data ranges
for each outcome
  build rules using data ranges of each neuron
prune insignificant rules
update data ranges in rule conditions analyzing error
```

***black box auditing***

if $((data(I_1) \geq L_{13} \wedge data(I_1) \leq U_{13}) \wedge (data(I_2) \geq L_{23} \wedge data(I_2) \leq U_{23}) \wedge (data(I_3) \geq L_{33} \wedge data(I_3) \leq U_{33}))$ then class $= C_3$

else

if $((data(I_1) \geq L_{11} \wedge data(I_1) \leq U_{11}) \wedge (data(I_3) \geq L_{31} \wedge data(I_3) \leq U_{31}))$ then class $= C_1$

else

class $= C_2$

- M. Gethsiyal Augasta and T. Kathirvalavakumar. 2012. ***Reverse engineering the neural networks for rule extraction in classification problems***. NPL.

| Name | Ref. | Authors | Year | Explanator | Black Box | Data Type | General | Random | Examples | Code | Dataset |
|---|---|---|---|---|---|---|---|---|---|---|---|
| – | [134] | Xu et al. | 2015 | SM | DNN | IMG | | | ✓ | ✓ | ✓ |
| – | [30] | Fong et al. | 2017 | SM | DNN | IMG | | | ✓ | | |
| CAM | [139] | Zhou et al. | 2016 | SM | DNN | IMG | | | ✓ | ✓ | ✓ |
| Grad-CAM | [106] | Selvaraju et al. | 2016 | SM | DNN | IMG | | | ✓ | ✓ | ✓ |
| – | [109] | Simonian et al. | 2013 | SM | DNN | IMG | | | ✓ | | ✓ |
| PWD | [7] | Bach et al. | 2015 | SM | DNN | IMG | | | ✓ | | ✓ |
| – | [113] | Sturm et al. | 2016 | SM | DNN | IMG | | | ✓ | | ✓ |
| DTD | [78] | Montavon et al. | 2017 | SM | DNN | IMG | | | ✓ | | ✓ |
| DeapLIFT | [107] | Shrikumar et al. | 2017 | FI | DNN | ANY | | | ✓ | ✓ | |
| CP | [64] | Landecker et al. | 2013 | SM | NN | IMG | | | ✓ | | |
| – | [143] | Zintgraf et al. | 2017 | SM | DNN | IMG | | | ✓ | ✓ | ✓ |
| VBP | [11] | Bojarski et al. | 2016 | SM | DNN | IMG | | | ✓ | | ✓ |
| – | [65] | Lei et al. | 2016 | SM | DNN | TXT | | | ✓ | | ✓ |
| ExplainD | [89] | Poulin et al. | 2006 | FI | SVM | TAB | | ✓ | ✓ | | |
| – | [29] | Strumbelj et al. | 2010 | FI | AGN | TAB | ✓ | ✓ | ✓ | | ✓ |

# Solving The Outcome Explanation Problem

# Local Model Explainers

- Explanator: SM
  - Black Box: DNN, NN
  - Data Type: IMG

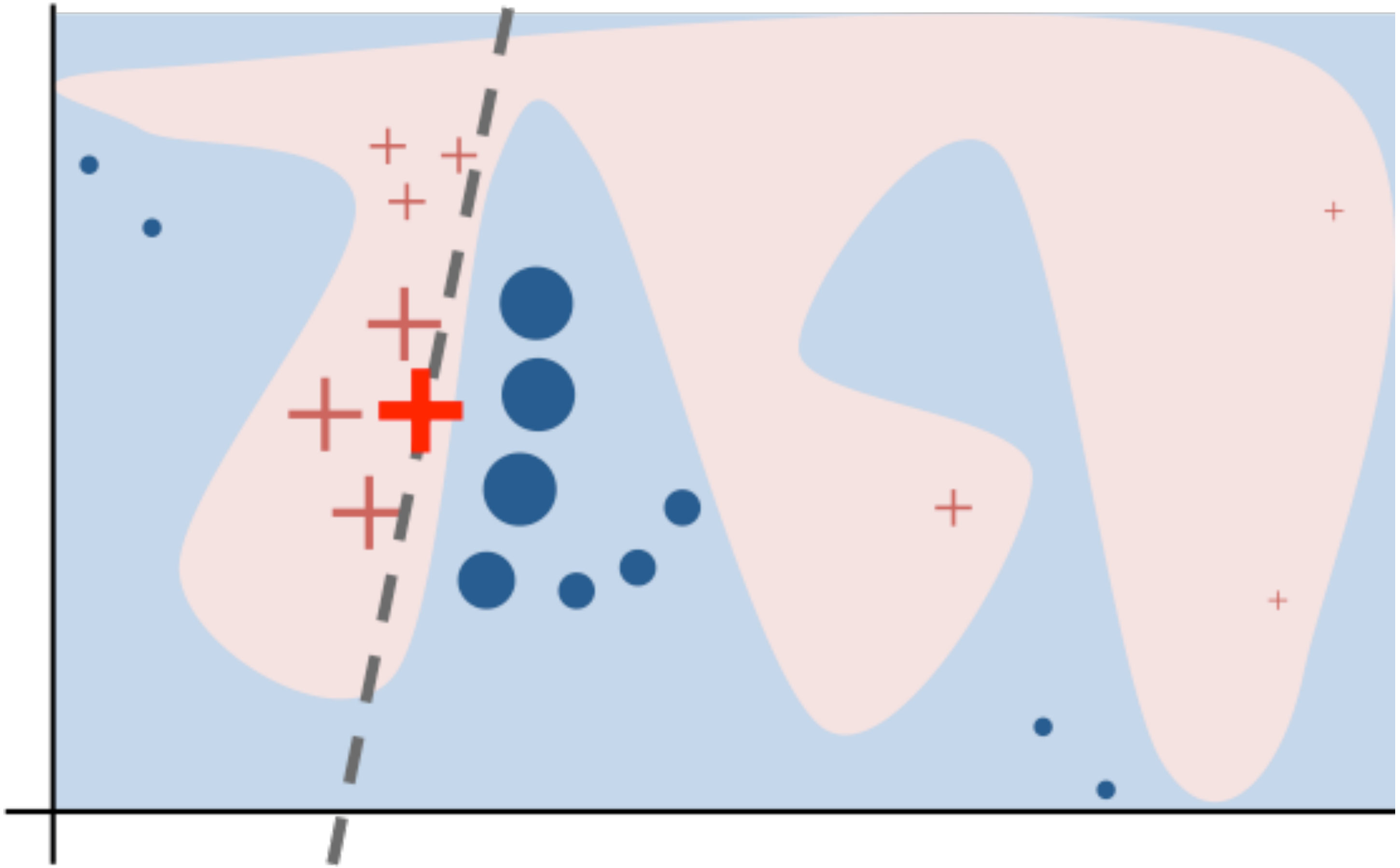
- Explanator: FI
  - Black Box: DNN, SVM
  - Data Type: ANY

- Explanator: DT
  - Black Box: ANY
  - Data Type: TAB

$R_1$: IF(Outlook = Sunny) AND (Windy= False) THEN Play=Yes

# Local Explanation

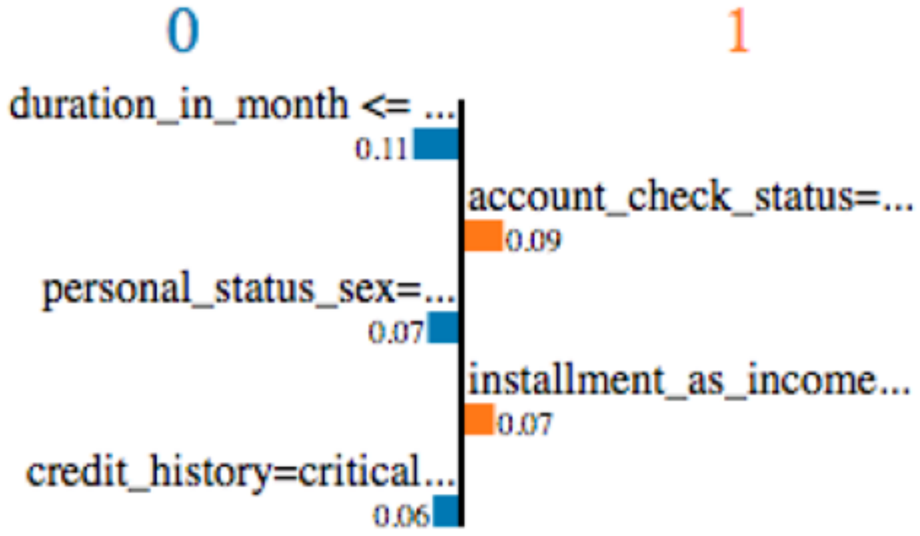
- The overall decision boundary is complex
- In the neighborhood of a single decision, the boundary is simple
- A single decision can be explained by auditing the black box around the given instance and learning a ***local*** decision.



# LIME – FI, AGN, "ANY"

```
Z = {}
x instance to explain
x' = real2interpretable(x)
for i in {1, 2, …, N}
       z_i = sample_around(x')
       z = interpretabel2real(z_i)
       Z = Z ∪ {<z_i, b(z_i), d(x, z)>}
w = solve_Lasso(Z, k)
return w
```

***black box auditing***




- Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. Why should i trust you?: Explaining the predictions of any classifier. KDD.


https://xaitutorial2019.github.io/

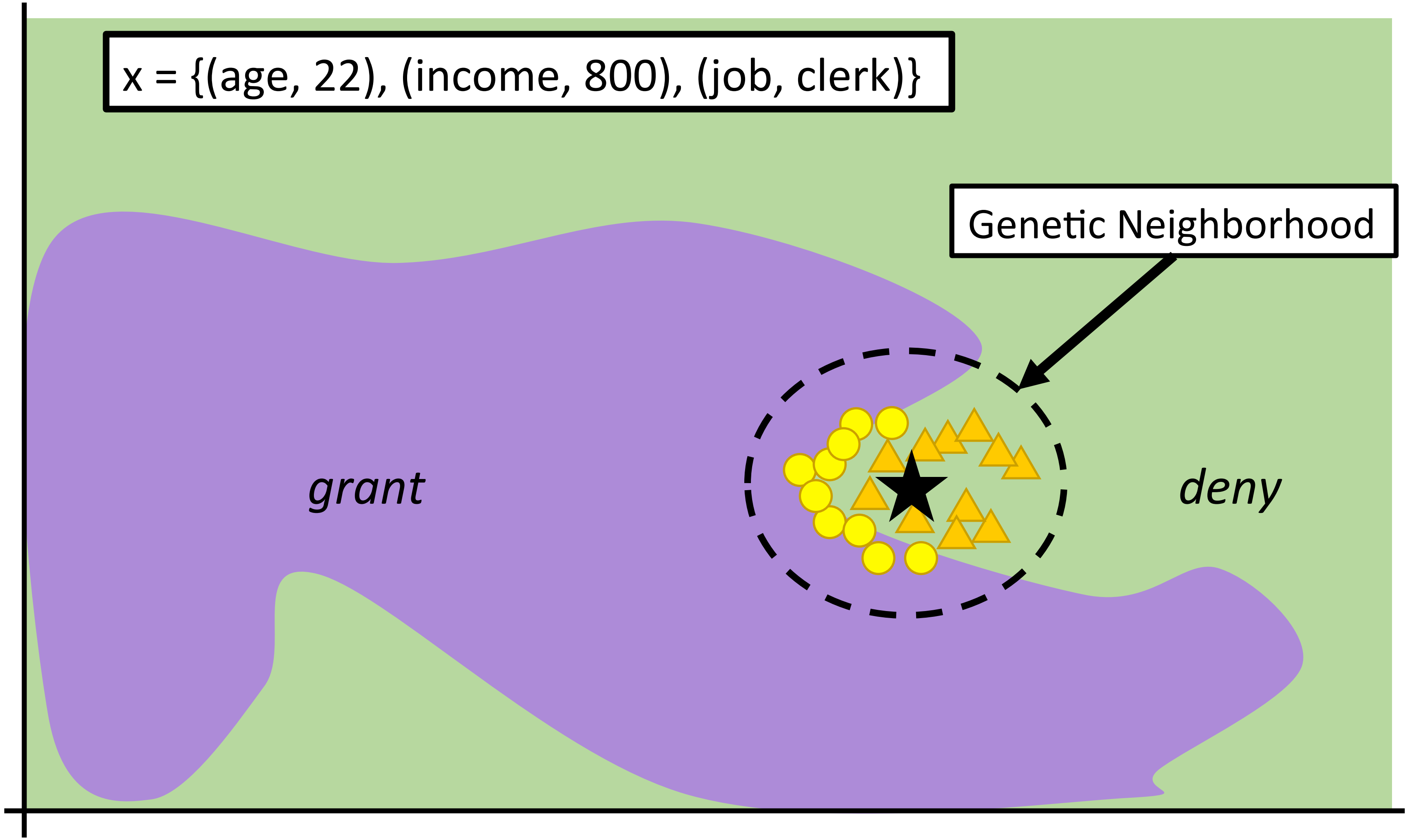
# LORE – DR, AGN, TAB

```
x instance to explain
Z_= = geneticNeighborhood(x, fitness_=, N/2)
Z_≠ = geneticNeighborhood(x, fitness_≠, N/2)
Z = Z_= ∪ Z_≠
c = buildTree(Z, b(Z))
r = (p -> y) = extractRule(c, x)
φ = extractCounterfactual(c, r, x)
return e = <r, φ>
```

***black box auditing***

Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Dino Pedreschi, Franco Turini, and Fosca Giannotti. 2018. ***Local rule-based explanations of black box decision systems***. arXiv preprint arXiv:1805.10820

# LORE: Local Rule-Based Explanations




crossover

| | | | | |
|---|---|---|---|---|
| parent 1 | 25 | clerk | 10k | yes |
| parent 2 | 30 | other | 5k | no |
| | | ↓ | | |
| children 1 | 25 | other | 5k | yes |
| children 2 | 30 | clerk | 10k | no |

mutation

| | | | | |
|---|---|---|---|---|
| parent | 25 | clerk | 10k | yes |
| | ↓ | | ↓ | |
| children | 27 | clerk | 7k | yes |

Fitness Function evaluates which elements are the "best life forms", that is, most appropriate for the result.

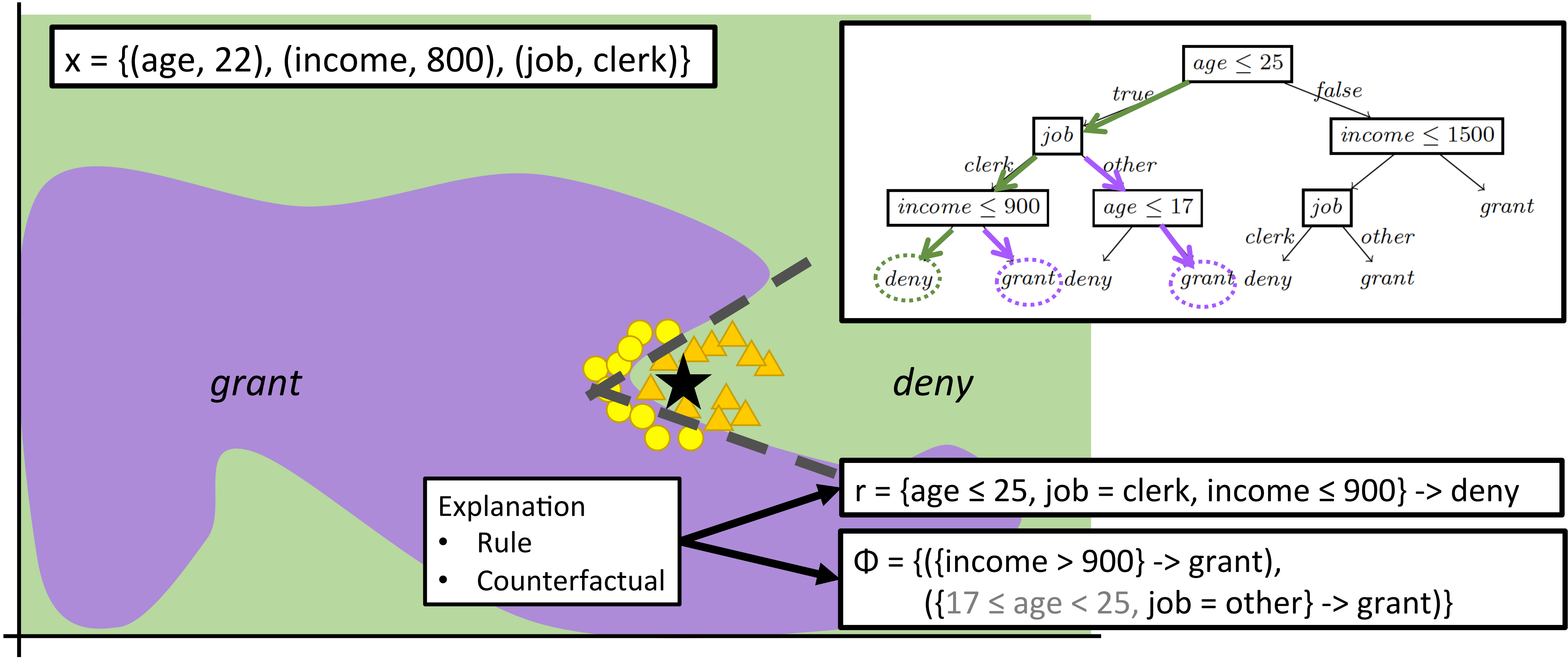
**fitness**

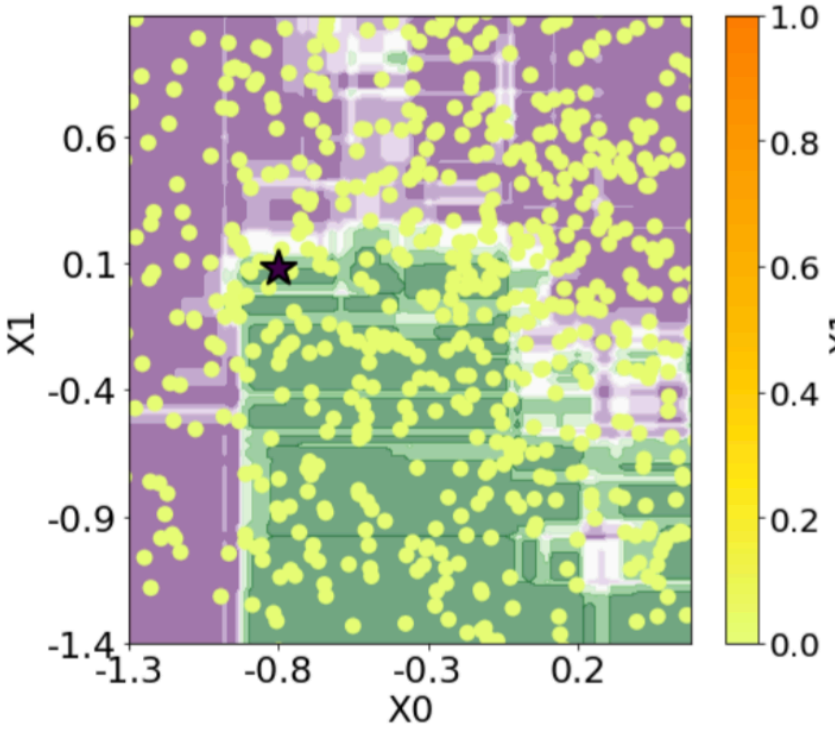
$$fitness_{=}^{x}(z) = I_{b(x)=b(z)} + (1 - d(x,z)) - I_{x=z}$$

$$fitness_{\neq}^{x}(z) = I_{b(x)\neq b(z)} + (1 - d(x,z)) - I_{x=z}$$

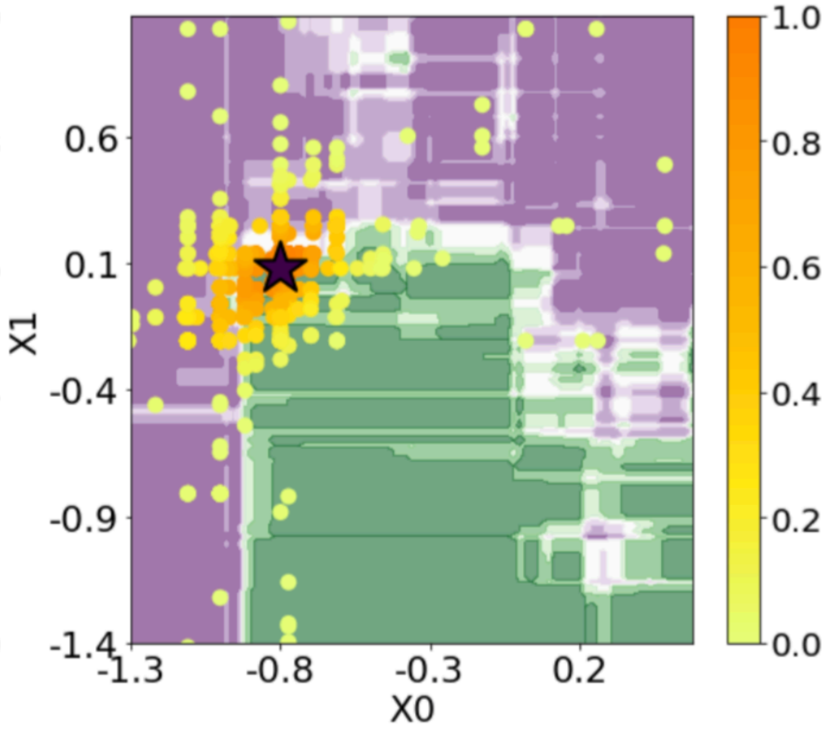
- Guidotti, R., Monreale, A., Ruggieri, S., Pedreschi, D., Turini, F., & Giannotti, F. (2018). ***Local Rule-Based Explanations of Black Box Decision Systems***. arXiv:1805.10820.

# Local Rule-Based Explanations

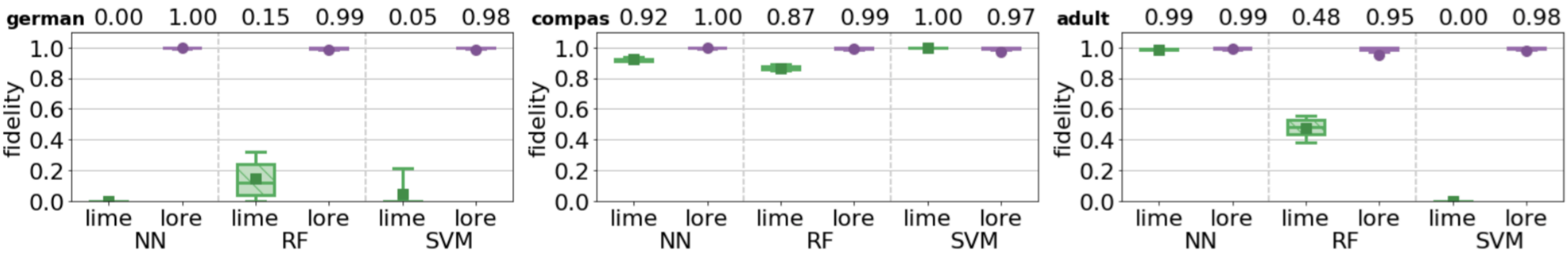




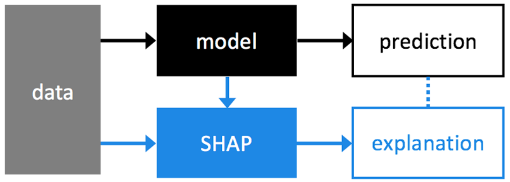

Random Neighborhood




Genetic Neighborhood



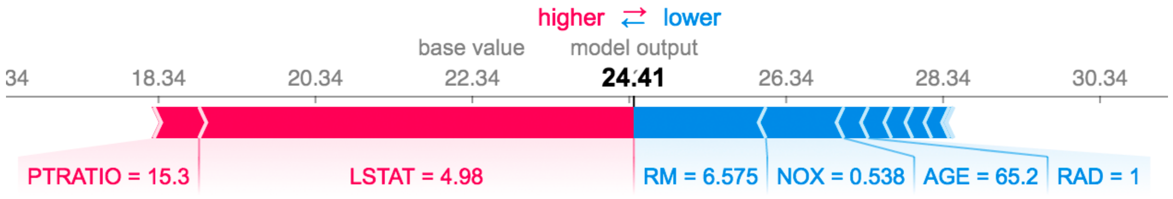
# SHAP (SHapley Additive exPlanations)

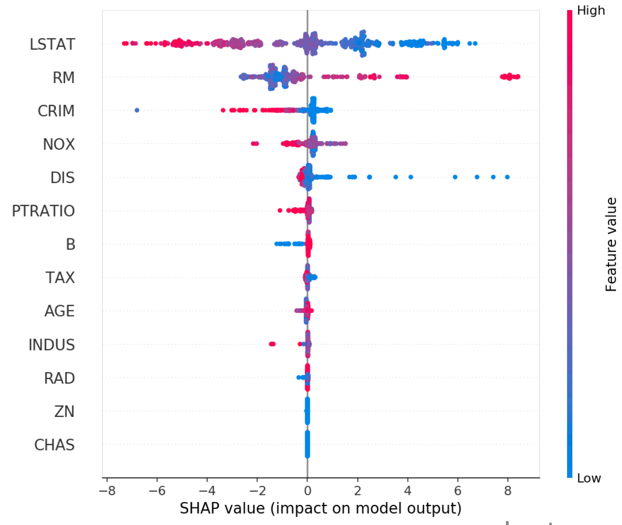



- SHAP assigns each feature an importance value for a particular prediction by means of an additive feature attribution method.
- It assigns an importance value to each feature that represents the effect on the model prediction of including that feature

$$g(z') = \phi_0 + \sum_{i=1}^{M} \phi_i z'_i,$$

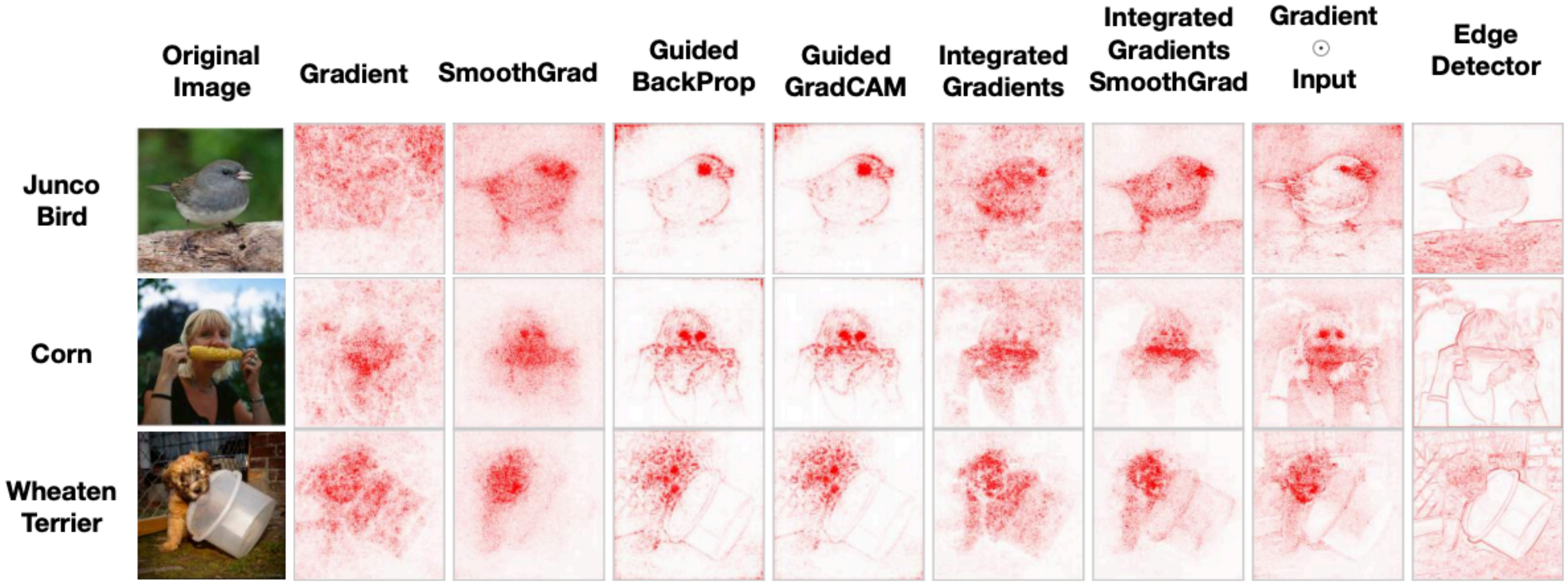
$$\phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!(|F| - |S| - 1)!}{|F|!} \left[ f_{S \cup \{i\}}(x_{S \cup \{i\}}) - f_S(x_S) \right]$$




- Lundberg, Scott M., and Su-In Lee. "A unified approach to interpreting model predictions." *Advances in Neural Information Processing Systems*. 2017.

# Saliency maps




Julius Adebayo, Justin Gilmer, Michael Christoph Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. 2018.

# Meaningful Perturbations – SM, DNN, IMG

01 x instance to explain

02 ***varying*** x into x′ maximizing b(x)~b(x′) ← ***black box auditing***

03 the variation runs replacing a region R of x with:

*constant value, noise, blurred image*

04 reformulation: find ***smallest*** R such that $b(x_R) \ll b(x)$


- Ruth Fong and Andrea Vedaldi. 2017. ***Interpretable explanations of black boxes by meaningful perturbation***. arXiv:1704.03296 (2017).

# Interpretable recommendations


L. Hu, S. Jian, L. Cao, and Q. Chen. Interpretable recommendation via attraction modeling: Learning multilevel attractiveness over multimodal movie contents. IJCAI-ECAI, 2018.

# Solving The Model Inspection Problem

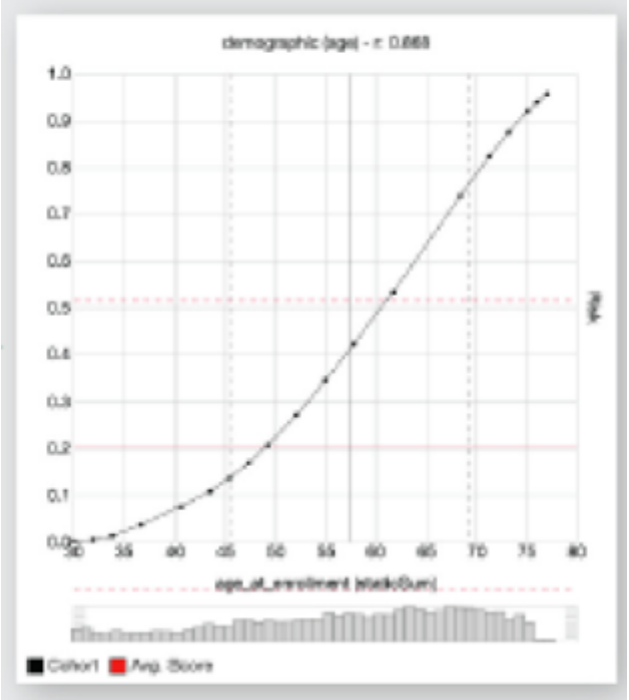
| Name | Ref. | Authors | Year | Explanator | Black Box | Data Type | General | Random | Examples | Code | Dataset |
|---|---|---|---|---|---|---|---|---|---|---|---|
| NID | [83] | Olden et al. | 2002 | SA | NN | TAB | | | ✓ | | |
| GDP | [8] | Baehrens | 2010 | SA | AGN | TAB | ✓ | | ✓ | | ✓ |
| QII | [24] | Datta et al | 2016 | SA | AGN | TAB | ✓ | | ✓ | | ✓ |
| IG | [115] | Sundararajan | 2017 | SA | DNN | ANY | | | ✓ | | ✓ |
| VEC | [18] | Cortez et al. | 2011 | SA | AGN | TAB | ✓ | | ✓ | | ✓ |
| VIN | [42] | Hooker | 2004 | PDP | AGN | TAB | ✓ | | ✓ | | ✓ |
| ICE | [35] | Goldstein et al. | 2015 | PDP | AGN | TAB | ✓ | | ✓ | ✓ | ✓ |
| Prospector | [55] | Krause et al. | 2016 | PDP | AGN | TAB | ✓ | | ✓ | | ✓ |
| Auditing | [2] | Adler et al. | 2016 | PDP | AGN | TAB | ✓ | | ✓ | ✓ | ✓ |
| OPIA | [1] | Adebayo et al. | 2016 | PDP | AGN | TAB | ✓ | | ✓ | | |
| – | [136] | Yosinski et al. | 2015 | AM | DNN | IMG | | | ✓ | | ✓ |
| IP | [108] | Shwartz et al. | 2017 | AM | DNN | TAB | | | ✓ | | |
| – | [137] | Zeiler et al. | 2014 | AM | DNN | IMG | | ✓ | | ✓ | |
| – | [112] | Springenberg et al. | 2014 | AM | DNN | IMG | | | ✓ | | ✓ |
| DGN-AM | [80] | Nguyen et al. | 2016 | AM | DNN | IMG | | | ✓ | ✓ | ✓ |

# Inspection Model Explainers

- Explanator: SA
  - Black Box: NN, DNN, AGN
  - Data Type: TAB

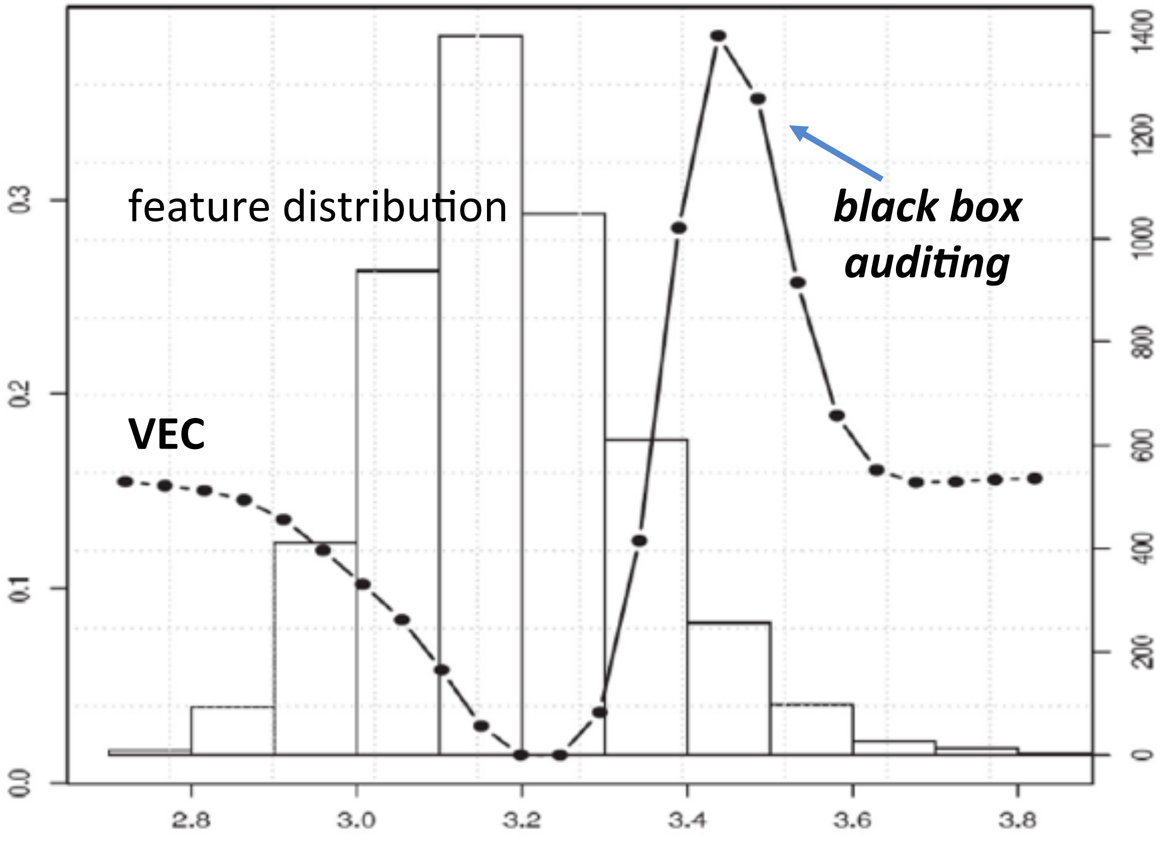
- Explanator: PDP
  - Black Box: AGN
  - Data Type: TAB

- Explanator: AM
  - Black Box: DNN
  - Data Type: IMG, TXT



# VEC – SA, AGN, TAB

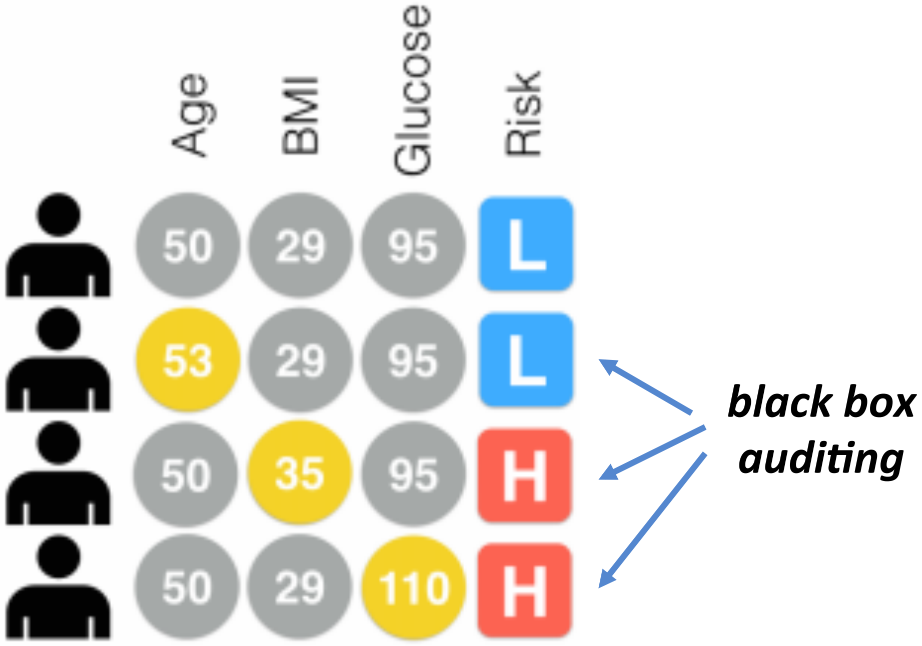
- Sensitivity measures are variables calculated as the range, gradient, variance of the prediction.
- The visualizations realized are barplots for the features importance, and ***Variable Effect Characteristic*** curve (VEC) plotting the input values versus the (average) outcome responses.

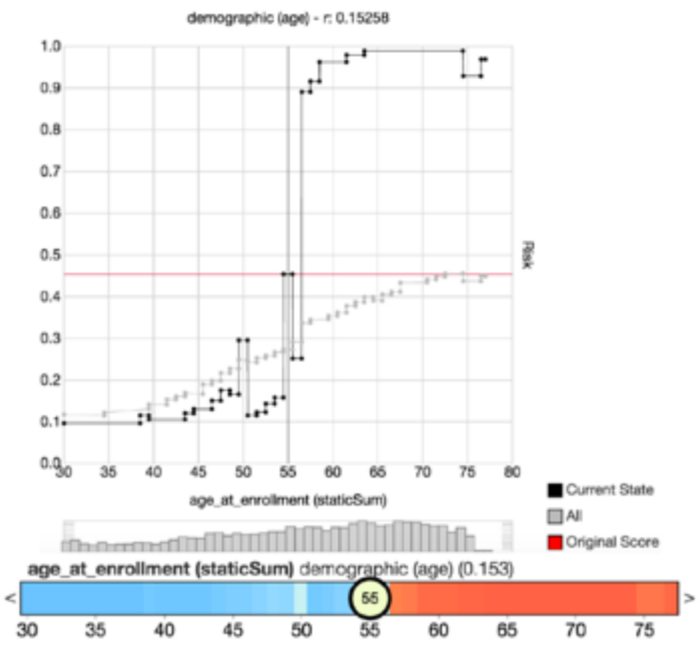



- Paulo Cortez and Mark J. Embrechts. 2011. ***Opening black box data mining models using sensitivity analysis***. CIDM.

# Prospector – PDP, AGN, TAB

- Introduce ***random perturbations*** on input values to understand to which extent every feature impact the prediction using PDPs.
- The input is changed ***one variable at a time***.




- Ruth Fong and Andrea Vedaldi. 2017. ***Interpretable explanations of black boxes by meaningful perturbation***. arXiv:1704.03296 (2017).

# Software disponibile

- LIME: https://github.com/marcotcr/lime
- MAPLE: https://github.com/GDPlumb/MAPLE
- SHAP: https://github.com/slundberg/shap
- ANCHOR: https://github.com/marcotcr/anchor
- LORE: https://github.com/riccotti/LORE
- https://ico.org.uk/media/about-the-ico/consultations/2616434/explaining-ai-decisions-part-1.pdf

# (Some) Software Resources

- **DeepExplain**: perturbation and gradient-based attribution methods for Deep Neural Networks interpretability. github.com/marcoancona/DeepExplain
- **iNNvestigate**: A toolbox to iNNvestigate neural networks' predictions. github.com/albermax/innvestigate
- **SHAP**: SHapley Additive exPlanations. github.com/slundberg/shap
- **ELI5**: A library for debugging/inspecting machine learning classifiers and explaining their predictions. github.com/TeamHG-Memex/eli5
- **Skater**: Python Library for Model Interpretation/Explanations. github.com/datascienceinc/Skater
- **Yellowbrick**: Visual analysis and diagnostic tools to facilitate machine learning model selection. github.com/DistrictDataLabs/yellowbrick
- **Lucid:** A collection of infrastructure and tools for research in neural network interpretability. github.com/tensorflow/lucid

# References


- Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). ***A survey of methods for explaining black box models***. *ACM Computing Surveys (CSUR)*, *51*(5), 93
- Finale Doshi-Velez and Been Kim. 2017. ***Towards a rigorous science of interpretable machine learning***. arXiv:1702.08608v2
- Alex A. Freitas. 2014. ***Comprehensible classification models: A position paper***. ACM SIGKDD Explor. Newslett.
- Andrea Romei and Salvatore Ruggieri. 2014. ***A multidisciplinary survey on discrimination analysis***. Knowl. Eng.
- Yousra Abdul Alsahib S. Aldeen, Mazleena Salleh, and Mohammad Abdur Razzaque. 2015. ***A comprehensive review on privacy preserving data mining***. SpringerPlus
- Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. ***Why should i trust you?: Explaining the predictions of any classifier***. KDD.
- Houtao Deng. 2014. ***Interpreting tree ensembles with intrees***. arXiv preprint arXiv:1408.5456.
- Mark Craven and JudeW. Shavlik. 1996. ***Extracting tree-structured representations of trained networks***. NIPS.

# References

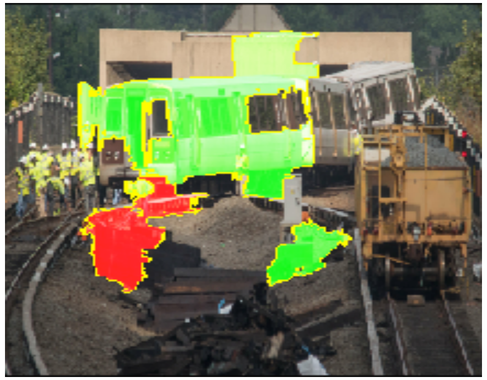

- M. Gethsiyal Augasta and T. Kathirvalavakumar. 2012. ***Reverse engineering the neural networks for rule extraction in classification problems***. NPL
- Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Dino Pedreschi, Franco Turini, and Fosca Giannotti. 2018. ***Local rule-based explanations of black box decision systems***. arXiv preprint arXiv:1805.10820
- Ruth Fong and Andrea Vedaldi. 2017. ***Interpretable explanations of black boxes by meaningful perturbation***. arXiv:1704.03296 (2017).
- Paulo Cortez and Mark J. Embrechts. 2011. ***Opening black box data mining models using sensitivity analysis***. CIDM.


- Ruth Fong and Andrea Vedaldi. 2017. ***Interpretable explanations of black boxes by meaningful perturbation***. arXiv:1704.03296 (2017).


- Xiaoxin Yin and Jiawei Han. 2003. ***CPAR: Classification based on predictive association rules***. SIAM, 331–335
- Angelino, E., Larus-Stone, N., Alabi, D., Seltzer, M., & Rudin, C. 2017. ***Learning certifiably optimal rule lists***. KDD.

# Applications

# Obstacle Identification Certification (Trust) - Transportation

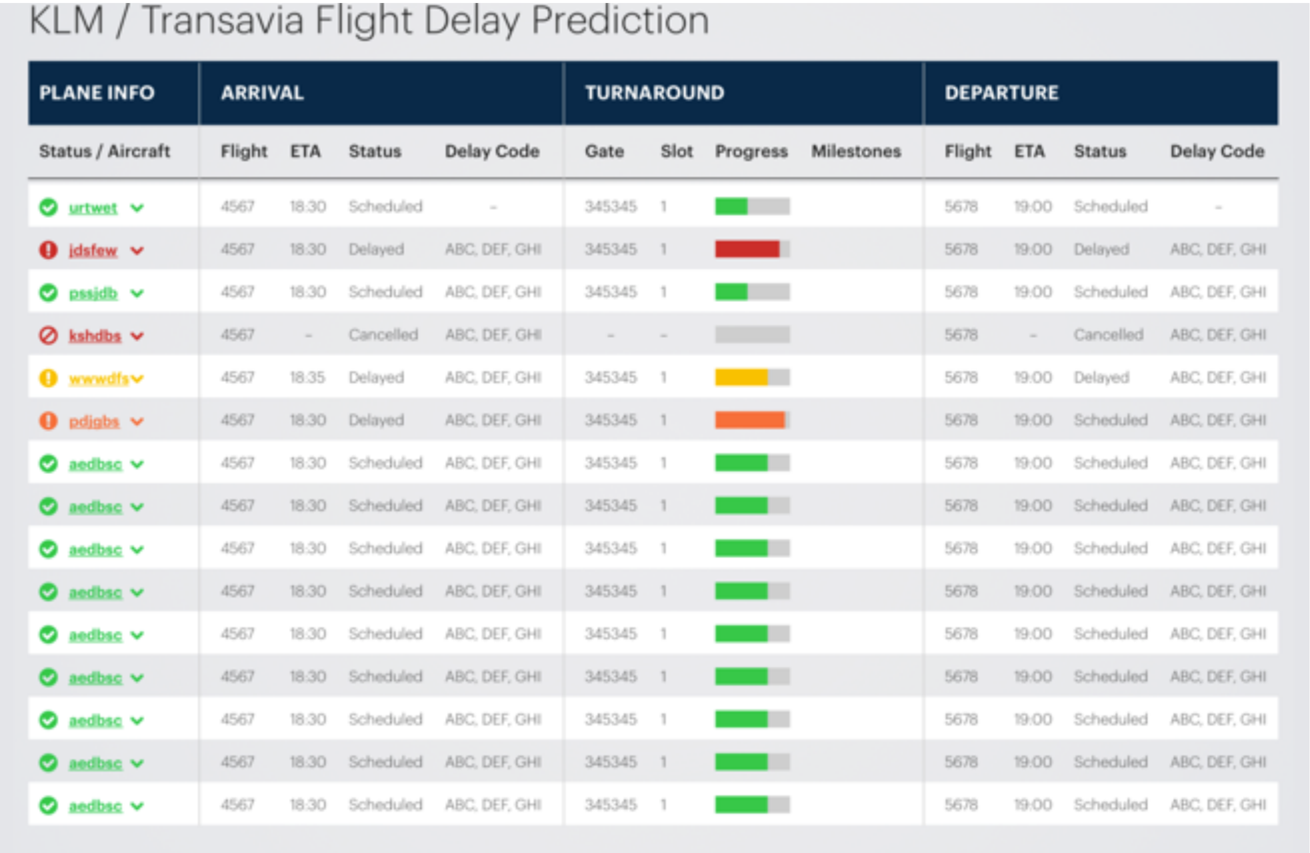


**Challenge:** Public transportation is getting more and more self-driving vehicles. Even if trains are getting more and more autonomous, the human stays in the loop for critical decision, for instance in case of obstacles. In case of obstacles trains are required to provide recommendation of action i.e., go on or go back to station. In such a case the human is required to validate the recommendation through an explanation exposed by the train or machine.

**AI Technology**: Integration of AI related technologies i.e., Machine Learning (Deep Learning / CNNs), and semantic segmentation.

**XAI Technology**: Deep learning and Epistemic uncertainty

# Explainable On-Time Performance - Transportation



KLM / Transavia Flight Delay Prediction

| PLANE INFO | ARRIVAL | | | | TURNAROUND | | | | DEPARTURE | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Status / Aircraft | Flight | ETA | Status | Delay Code | Gate | Slot | Progress | Milestones | Flight | ETA | Status | Delay Code |
| urtwet | 4567 | 18:30 | Scheduled | - | 345345 | 1 | | | 5678 | 19:00 | Scheduled | - |
| idsfew | 4567 | 18:30 | Delayed | ABC, DEF, GHI | 345345 | 1 | | | 5678 | 19:00 | Delayed | ABC, DEF, GHI |
| pssjdb | 4567 | 18:30 | Scheduled | ABC, DEF, GHI | 345345 | 1 | | | 5678 | 19:00 | Scheduled | ABC, DEF, GHI |
| kshdbs | 4567 | - | Cancelled | ABC, DEF, GHI | - | - | | | 5678 | - | Cancelled | ABC, DEF, GHI |
| wwwdfs | 4567 | 18:35 | Delayed | ABC, DEF, GHI | 345345 | 1 | | | 5678 | 19:00 | Delayed | ABC, DEF, GHI |
| pdjgbs | 4567 | 18:30 | Delayed | ABC, DEF, GHI | 345345 | 1 | | | 5678 | 19:00 | Scheduled | ABC, DEF, GHI |
| aedbsc | 4567 | 18:30 | Scheduled | ABC, DEF, GHI | 345345 | 1 | | | 5678 | 19:00 | Scheduled | ABC, DEF, GHI |
| aedbsc | 4567 | 18:30 | Scheduled | ABC, DEF, GHI | 345345 | 1 | | | 5678 | 19:00 | Scheduled | ABC, DEF, GHI |
| aedbsc | 4567 | 18:30 | Scheduled | ABC, DEF, GHI | 345345 | 1 | | | 5678 | 19:00 | Scheduled | ABC, DEF, GHI |
| aedbsc | 4567 | 18:30 | Scheduled | ABC, DEF, GHI | 345345 | 1 | | | 5678 | 19:00 | Scheduled | ABC, DEF, GHI |
| aedbsc | 4567 | 18:30 | Scheduled | ABC, DEF, GHI | 345345 | 1 | | | 5678 | 19:00 | Scheduled | ABC, DEF, GHI |
| aedbsc | 4567 | 18:30 | Scheduled | ABC, DEF, GHI | 345345 | 1 | | | 5678 | 19:00 | Scheduled | ABC, DEF, GHI |
| aedbsc | 4567 | 18:30 | Scheduled | ABC, DEF, GHI | 345345 | 1 | | | 5678 | 19:00 | Scheduled | ABC, DEF, GHI |
| aedbsc | 4567 | 18:30 | Scheduled | ABC, DEF, GHI | 345345 | 1 | | | 5678 | 19:00 | Scheduled | ABC, DEF, GHI |
| aedbsc | 4567 | 18:30 | Scheduled | ABC, DEF, GHI | 345345 | 1 | | | 5678 | 19:00 | Scheduled | ABC, DEF, GHI |

**Challenge:** Globally 323,454 flights are delayed every year. Airline-caused delays totaled 20.2 million minutes last year, generating huge cost for the company. Existing in-house technique reaches 53% accuracy for **predicting flight delay**, does not provide any time estimation (in **minutes** as opposed to True/False) and is unable to capture the underlying reasons (explanation).

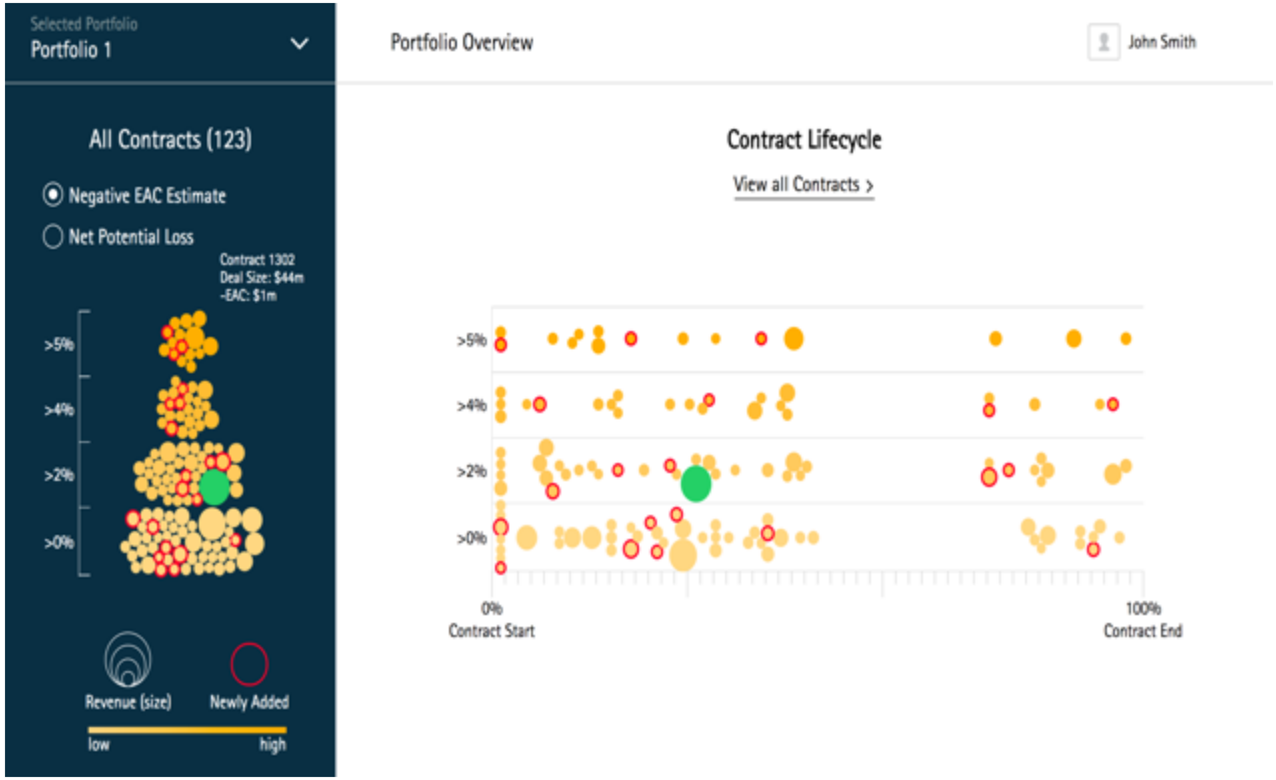
**AI Technology**: Integration of AI related technologies i.e., Machine Learning (Deep Learning / Recurrent neural Network), Reasoning (through semantics-augmented case-based reasoning) and Natural Language Processing for building a robust model which can (1) predict flight delays in minutes, (2) explain delays by comparing with historical cases.

**XAI Technology**: Knowledge graph embedded Sequence Learning using LSTMs

Jiaoyan Chen, Freddy Lécué, Jeff Z. Pan, Ian Horrocks, Huajun Chen: Knowledge-Based Transfer Learning Explanation. KR 2018: 349-358

Nicholas McCarthy, Mohammad Karzand, Freddy Lecue: Amsterdam to Dublin Eventually Delayed? LSTM and Transfer Learning for Predicting Delays of Low Cost Airlines: AAAI 2019

# Explainable Risk Management - Finance



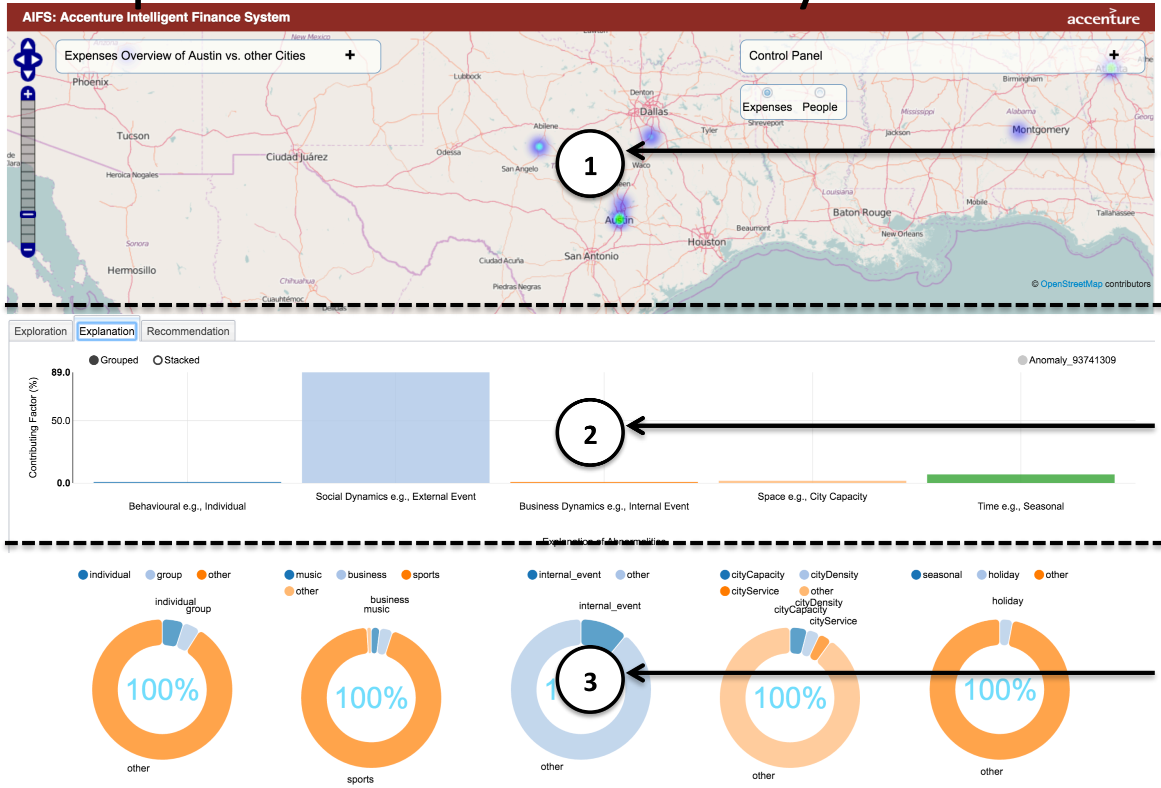

Jiewen Wu, Freddy Lécué, Christophe Guéret, Jer Hayes, Sara van de Moosdijk, Gemma Gallagher, Peter McCanney, Eugene Eichelberger: Personalizing Actions in Context for Risk Management Using Semantic Web Technologies. International Semantic Web Conference (2) 2017: 367-383

**Challenge:** Accenture is managing every year more than 80,000 opportunities and 35,000 contracts with an expected revenue of $34.1 billion. Revenue expectation does not meet estimation due to the complexity and risks of critical contracts. This is, in part, due to the (1) large volume of projects to assess and control, and (2) the existing non-systematic assessment process.

**AI Technology**: Integration of AI technologies i.e., Machine Learning, Reasoning, Natural Language Processing for building a robust model which can (1) predict revenue loss, (2) recommend corrective actions, and (3) explain why such actions might have a positive impact.

**XAI Technology:** Knowledge graph embedded Random Forrest

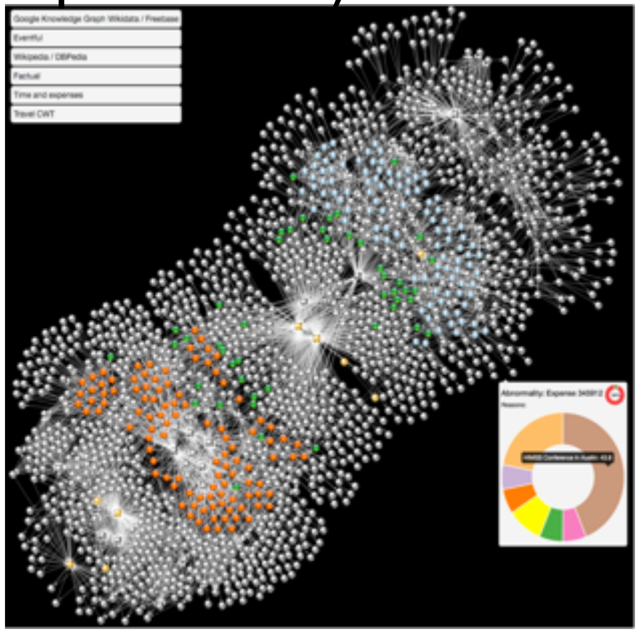
# Explainable anomaly detection – Finance (Compliance)




Data analysis for spatial interpretation of abnormalities: abnormal expenses

Semantic explanation (structured in classes: fraud, events, seasonal) of abnormalities

Detailed semantic explanation (structured in sub classes e.g. categories for events)



Freddy Lécué, Jiewen Wu: Explaining and predicting abnormal expenses at large scale using knowledge graph based reasoning. J. Web Sem. 44: 89-103 (2017)

**Challenge:** Predicting and explaining abnormally employee expenses (as high accommodation price in 1000+ cities).

**AI Technology:** Various techniques have been matured over the last two decades to achieve excellent results. However most methods address the problem from a statistic and pure data-centric angle, which in turn limit any interpretation. We elaborated a web application running live with real data from (i) travel and expenses from Accenture, (ii) external data from third party such as Google Knowledge Graph, DBPedia (relational DataBase version of Wikipedia) and social events from Eventful, for explaining abnormalities.

**XAI Technology:** Knowledge graph embedded Ensemble Learning

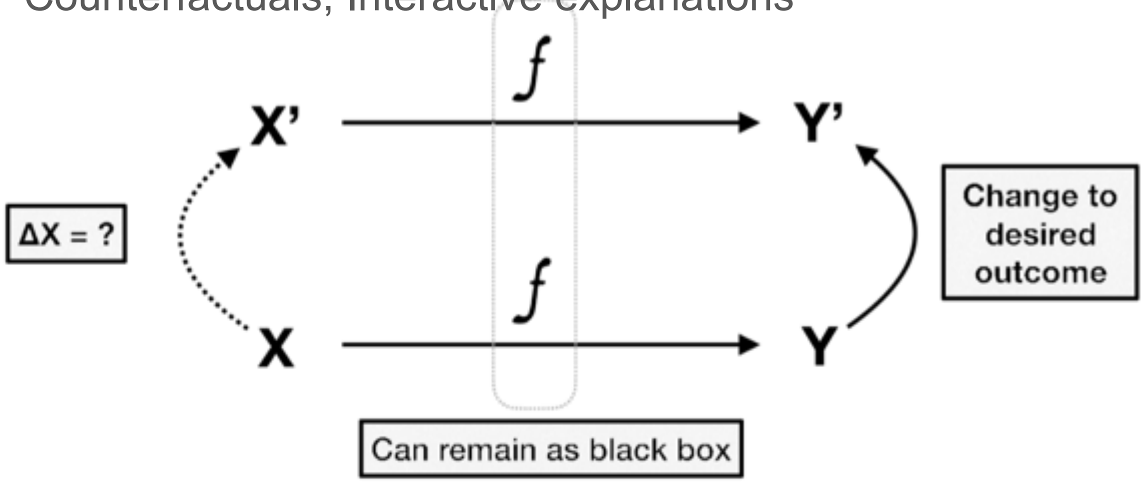
# Counterfactual Explanations for Credit Decisions

- Local, post-hoc, contrastive explanations of black-box classifiers
- **Required minimum change in input vector to flip the decision of the classifier.**
- Interactive Contrastive Explanations

**Challenge:** We predict loan applications with off-the-shelf, interchangeable black-box estimators, and we explain their predictions with counterfactual explanations. In counterfactual explanations the model itself remains a black box; it is only through changing inputs and outputs that an explanation is obtained.

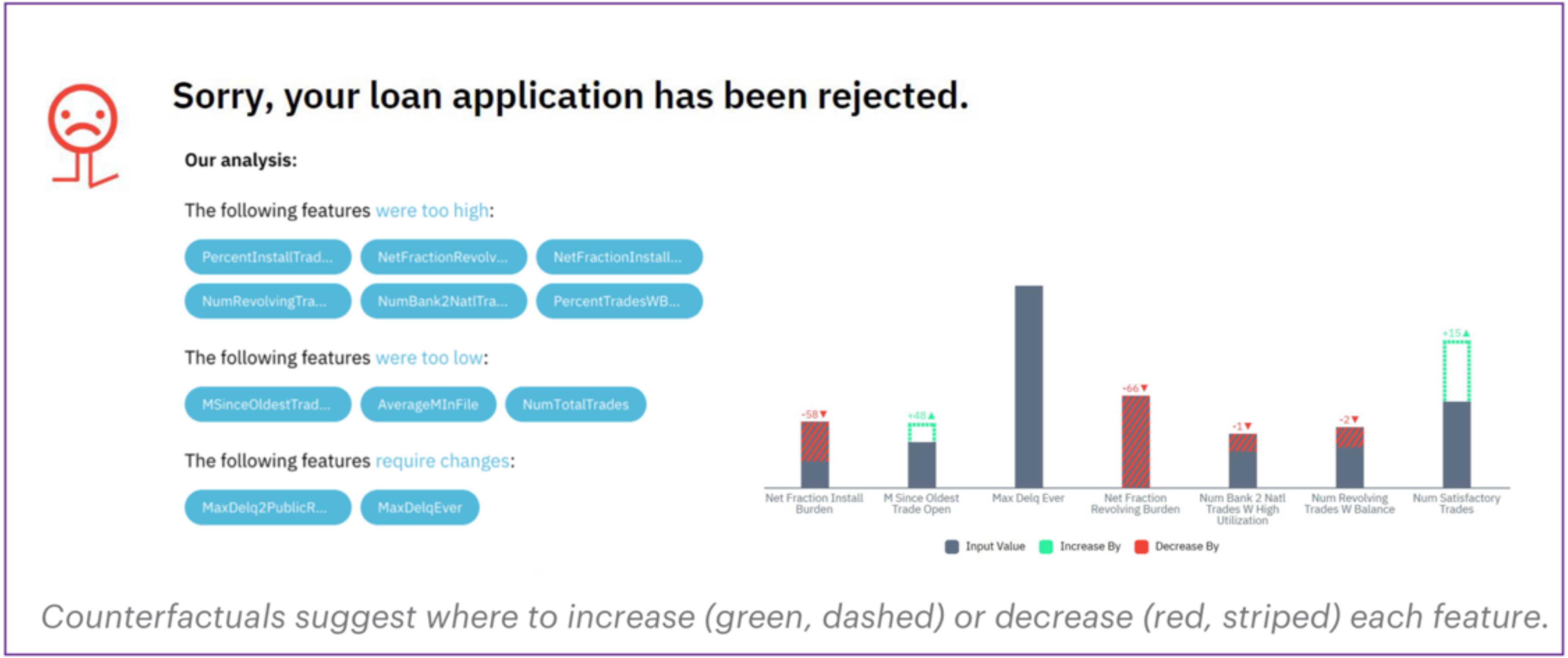
**AI Technology**: Supervised learning, binary classification.

**XAI Technology:** Post-hoc explanation, Local explanation, Counterfactuals, Interactive explanations

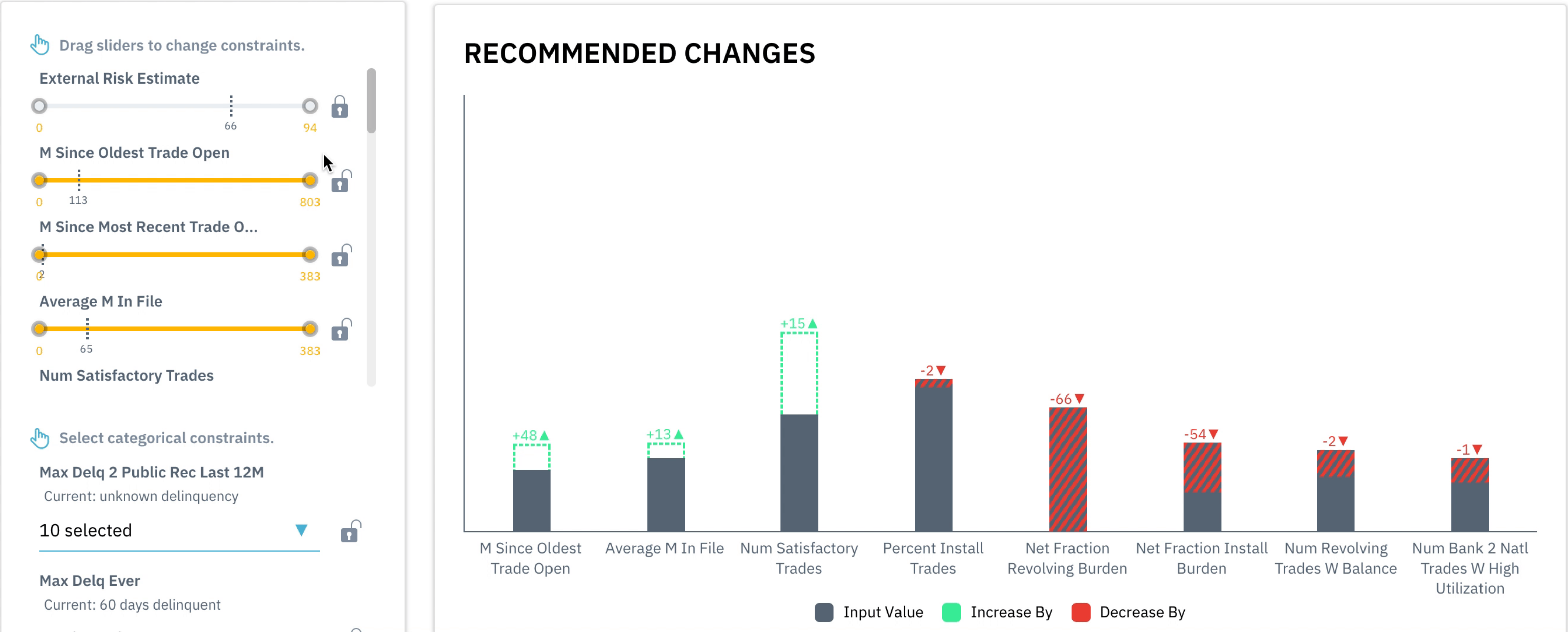



Rory Mc Grath, Luca Costabello, Chan Le Van, Paul Sweeney, Farbod Kamiab, Zhao Shen, Freddy Lécué: Interpretable Credit Application Predictions With Counterfactual Explanations. FEAP-AI4fin workshop, NeurIPS, 2018.

# Counterfactual Explanations for Credit Decisions



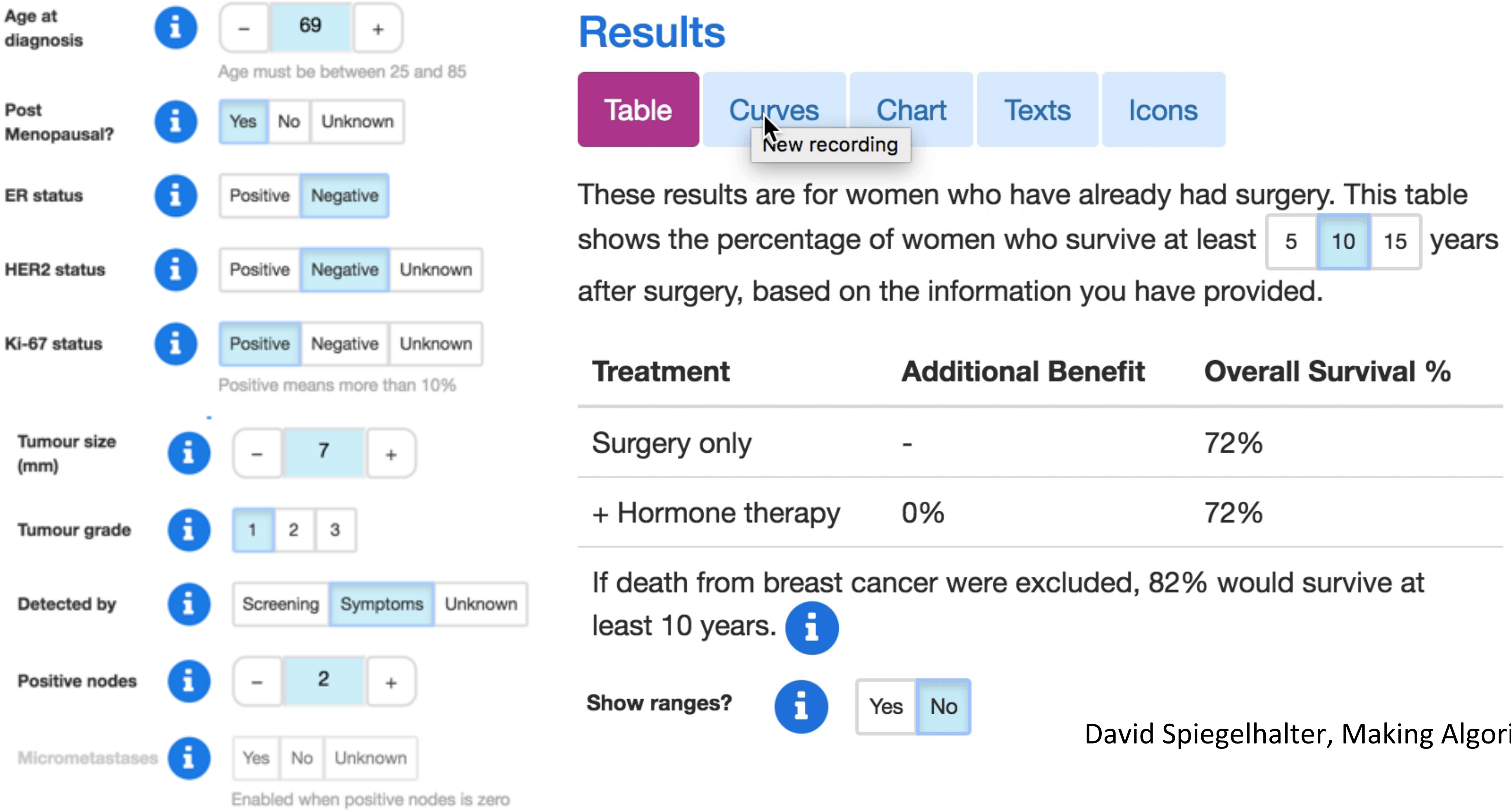

Rory Mc Grath, Luca Costabello, Chan Le Van, Paul Sweeney, Farbod Kamiab, Zhao Shen, Freddy Lécué: Interpretable Credit Application Predictions With Counterfactual Explanations. FEAP-AI4fin workshop, NeurIPS, 2018.




Rory Mc Grath, Luca Costabello, Chan Le Van, Paul Sweeney, Farbod Kamiab, Zhao Shen, Freddy Lécué: Interpretable Credit Application Predictions With Counterfactual Explanations. FEAP-AI4fin workshop, NeurIPS, 2018.

# Breast Cancer Survival Rate Prediction

predict breast cancer




**Challenge:** Predict is an online tool that helps patients and clinicians see how different treatments for early invasive breast cancer might improve survival rates after surgery.

**AI Technology**: competing risk analysis

**XAI Technology:** Interactive explanations, Multiple representations.

David Spiegelhalter, Making Algorithms trustworthy, NeurIPS 2018 Keynote

predict.nhs.uk/tool

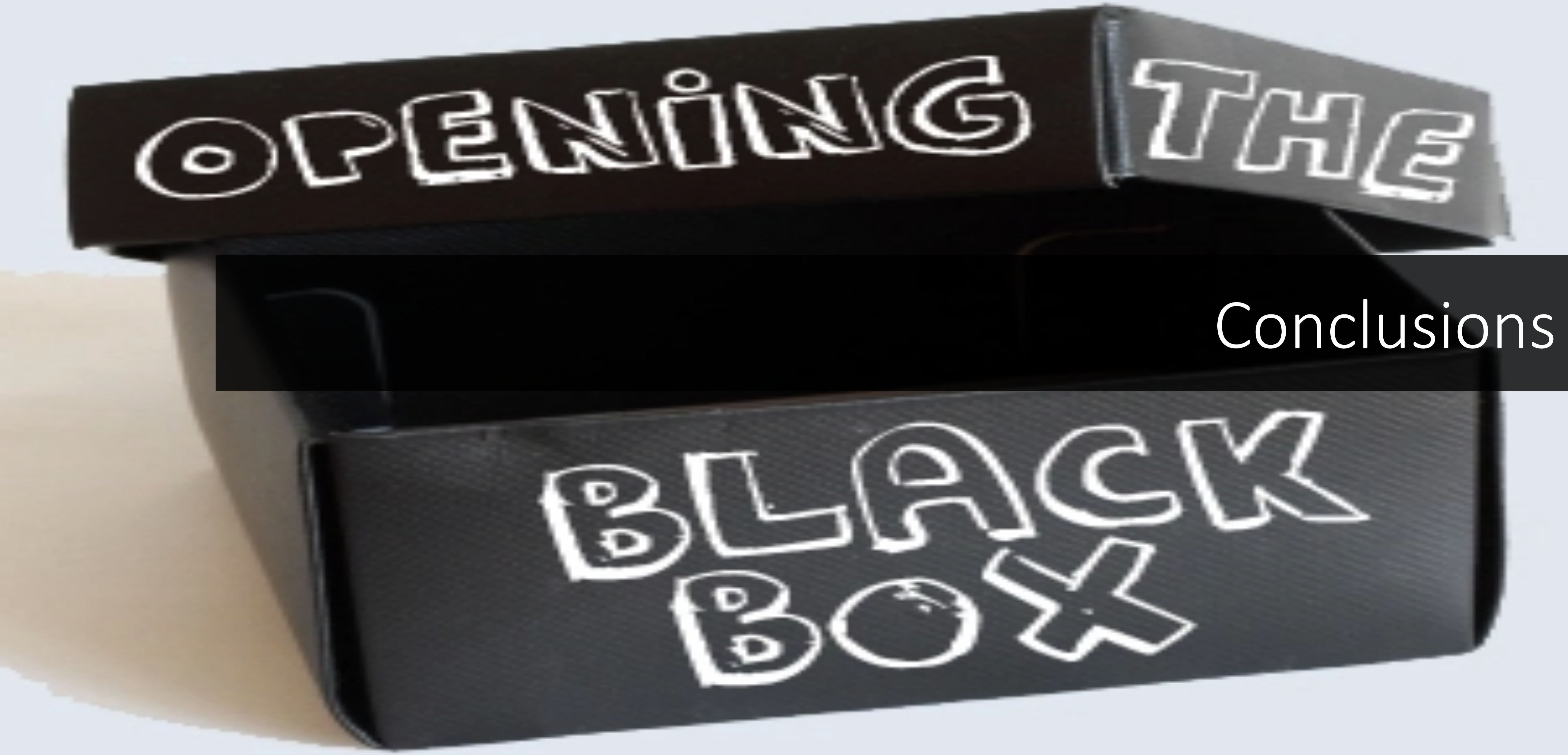
# Reasoning on Local Explanations of Classifications Operated by Black Box Models

- DIVA (Fraud Detection IVA) dataset from Agenzia delle Entrate containing about 34 milions IVA declarations and 123 features.
- 92.09% of the instances classified with label '3' by the KDD-Lab classifier are classified with the same instance and with an explanation by LORE.

| **Explanation** |
|---|
| VAL_ALIQ_MEDIA_ACQ>19.99, cod_uff_prov_gen=PR, IMP_V_AGG_IVA<=40264.00, VAR_DETRAZIONE>-334159.94 |
| VAL_ALIQ_MEDIA_ACQ>19.97, VAL_ALIQ_M_VOL_IMP>19.98, PESO_ADESIONE<=4.71, COD_MOD_DICH=6, RIMB_NON_CONC>-17351.76, MAG_IMP_RIT_ACC>-12519.81 |
| VAL_ALIQ_MEDIA_ACQ>19.87, VAL_ALIQ_MEDIA_VOL>19.01, IMP_IVA_DEB>2373859.00, DUR_P_PIVA_MM!=116, IMP_BEN_AMM<=2629.50 |

| **Jaccard** | **Avg DT len** | **Avg len** |
|---|---|---|
| 0.321 | 4.948 | 3.912 |

- Master Degree Thesis Leonardo Di Sarli, 2019




# Conclusions

# Guidance - Part 1 The basics of explaining AI

- https://ico.org.uk/media/about-the-ico/consultations/2616434/explaining-ai-decisions-part-1.pdf
- **Rationale explanation**: the reasons that led to a decision, delivered in an accessible and non-technical way.
- **Responsibility explanation**: who is involved in the development, management and implementation of an AI system, and who to contact for a human review of a decision.
- **Data explanation**: what data has been used in a particular decision and how; what data has been used to train and test the AI model and how.
- **Fairness explanation**: steps taken across the design and implementation of an AI system to ensure that the decisions it supports are generally unbiased and fair, and whether or not an individual has been treated equitably.
- **Safety and performance explanation**: steps taken across the design and implementation of an AI system to maximise the accuracy, reliability, security and robustness of its decisions and behaviours.
- **Impact explanation**: the impact that the use of an AI system and its decisions has or may have on an individual, and on wider society.

# Check -list

- We have identified everyone involved in the decision-making pipeline and where they are responsible for providing an explanation of the AI system.
- We have ensured that different actors along the decision-making pipeline, particularly those in AI development teams, those giving explanations to decision recipients, and our DPO and compliance teams are able to carry out their role in producing and delivering explanations.
- Where we are buying the AI system from a third party, we know we have the primarily responsibility for ensuring that the AI system is capable of producing explanations.