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

Motivating Examples

- Criminal Justice
  - People wrongly denied
  - Recidivism prediction
  - Unfair Police dispatch
- Finance:
  - Credit scoring, loan approval
  - Insurance quotes
- Healthcare
  - AI as 3rd-party actor in physician - patient relationship
  - Learning must be done with available data: cannot randomize cares given to patients!
  - Must validate models before use.

Opinion

The New York Times

OP-ED CONTRIBUTOR

When a Computer Program Keeps You in Jail

The Big Read Artificial intelligence + Add to myFT

Insurance: Robots learn the business of covering risk

Stanford MEDICINE | News Center

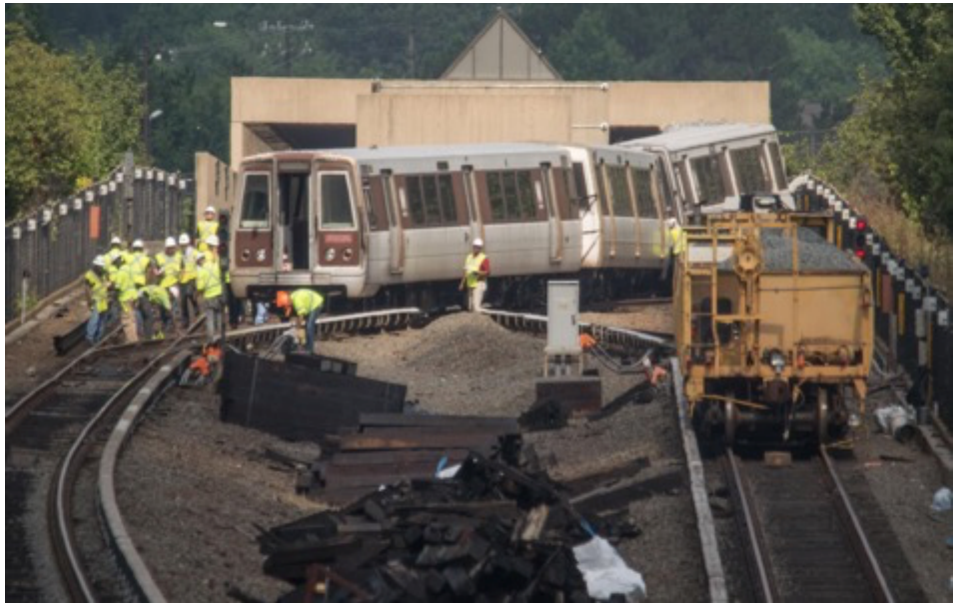
Email Tweet

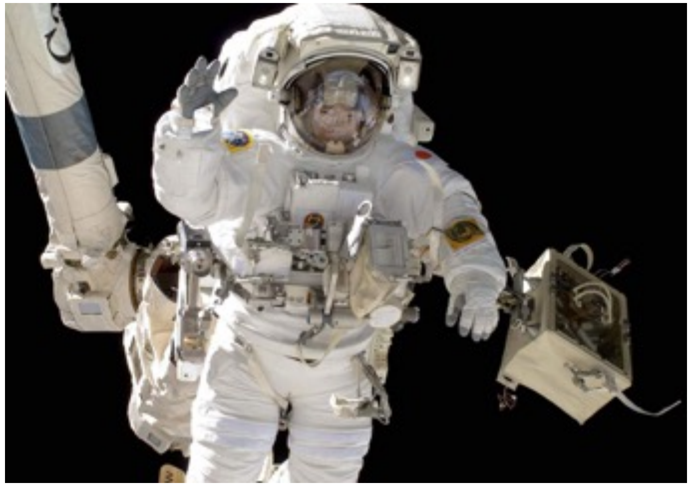
Researchers say use of artificial intelligence in medicine raises ethical questions

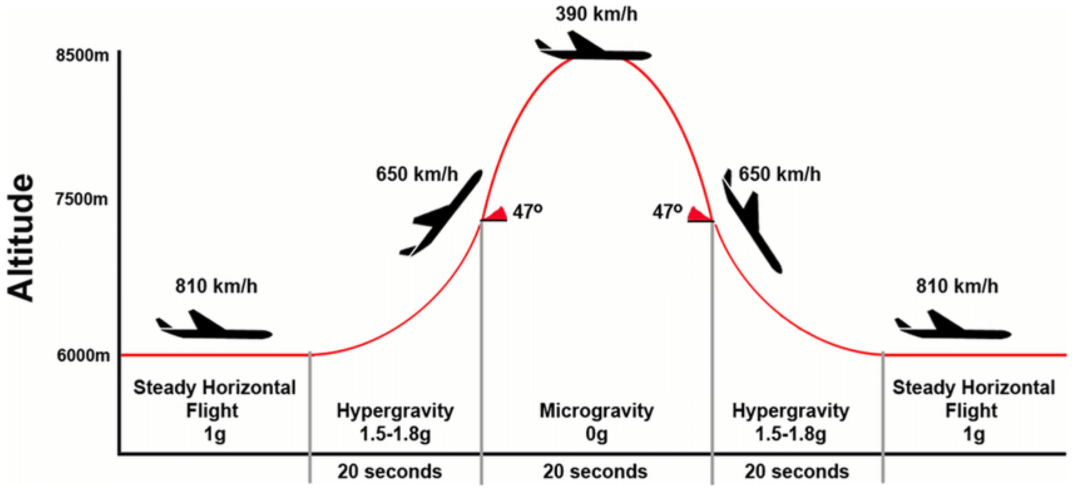
In a perspective piece, Stanford researchers discuss the ethical implications of using machine-learning tools in making health care decisions for patients.

Motivation (4)

- Critical Systems

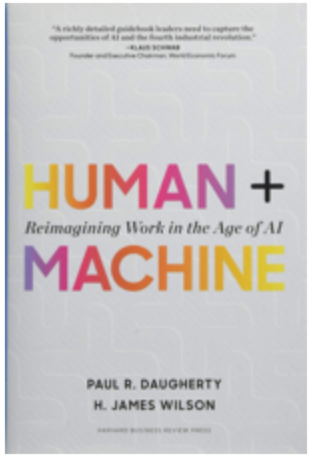






[Caruana et al. 2015, Holzinger et al. 2017, Magnus et al. 2018]

The Need for Explanation



- **Critical systems / Decisive moments**
- Human factor:
  - Human decision-making affected by **greed, prejudice, fatigue, poor scalability**.
  - **Bias**
- Algorithmic decision-making on the rise.
  - More objective than humans?
  - Potentially discriminative
  - Opaque
  - Information and power asymmetry
- High-stakes scenarios = **ethical** problems!

[Lepri et al. 2018]

Right of Explanation

Since 25 May 2018, GDPR establishes a right for all individuals to obtain “meaningful explanations of the logic involved” when “automated (algorithmic) individual decision-making”, including profiling, takes place.

Ethical principles for trustworthy AI

respect for human autonomy
  self-determination
  no-coercion
  no-manipulation

prevention of harm
  safe and secure

fairness
  no-discrimination (no-bias)

explicability
  User trust and transparency
  intelligibility "how does it work?"
  accountability ("who is responsible for")

References

[Caruana et al. 2015] Caruana, Rich, et al. "Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission." Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2015.

[Gunning 2017] Gunning, David. "Explainable artificial intelligence (xai)." Defense Advanced Research Projects Agency (DARPA), nd Web (2017).

[Holzinger et al. 2017] Andreas Holzinger, Bernd Malle, Peter Kieseberg, Peter M. Roth, Heimo Mller, Robert Reihs, and Kurt Zatloukal. Towards the augmented pathologist: Challenges of explainable-ai in digital pathology. arXiv:1712.06657, 2017.

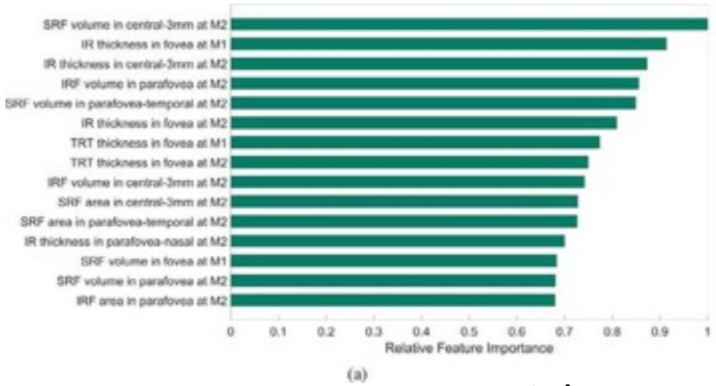
[Lepri et al. 2018] Lepri, Bruno, et al. "Fair, Transparent, and Accountable Algorithmic Decision-making Processes." Philosophy & Technology (2017): 1-17.

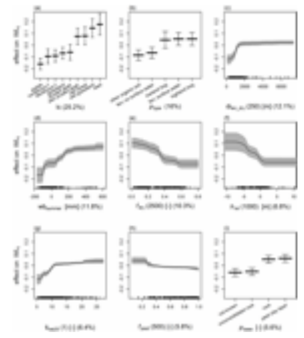
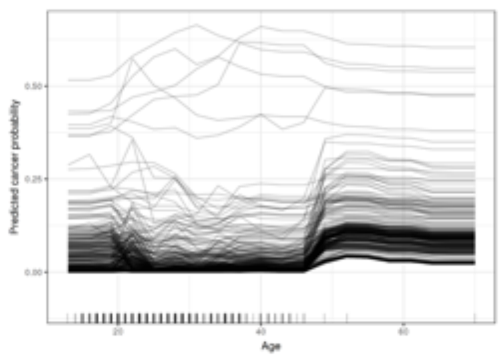
[Floridi et al**. 2019] Floridi,** Luciano and Josh Cowls "A Unified Framework of Five Principles for AI in Society". Harvard Data Science Review, 1, 2019

Explanation in AI

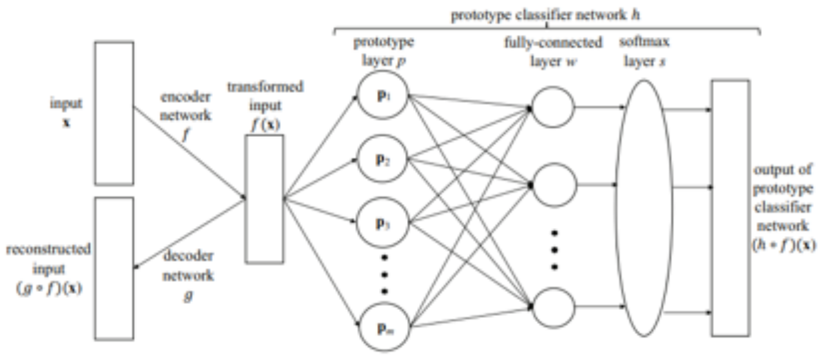
Explanation in different AI fields

- Machine Learning



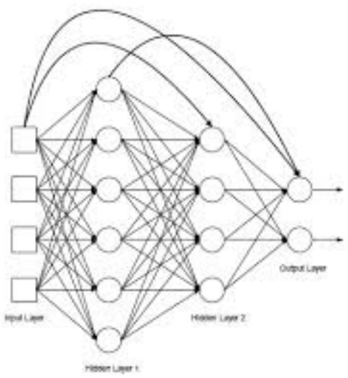
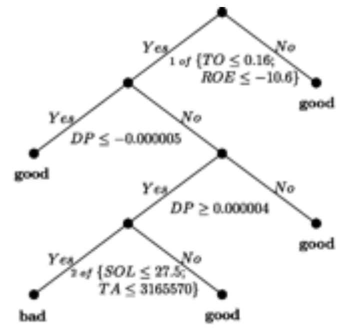



Feature Importance, Partial Dependence Plot, Individual Conditional Expectation



Auto-encoder

Oscar Li, Hao Liu, Chaofan Chen, Cynthia Rudin: Deep Learning for Case-Based Reasoning Through Prototypes: A Neural Network That Explains Its Predictions. AAAI 2018: 3530-3537

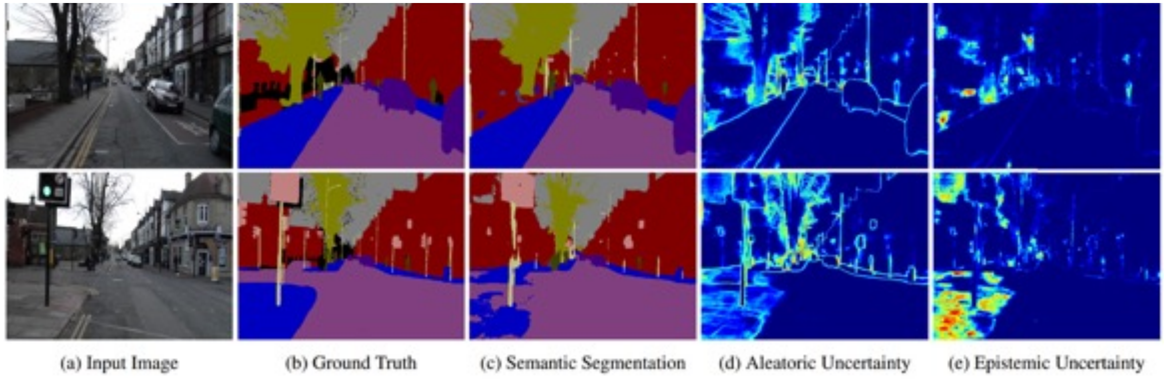



Surogate Model

Mark Craven, Jude W. Shavlik: Extracting Tree-Structured Representations of Trained Networks. NIPS 1995: 24-30

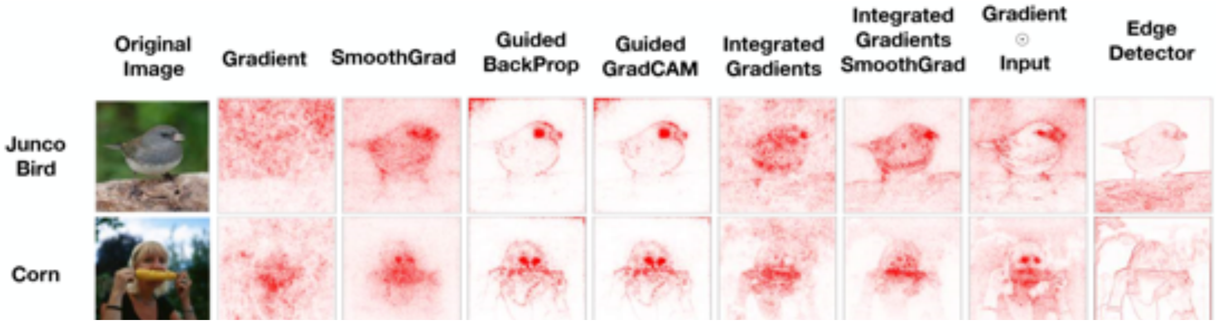
Explanation in different AI fields

- Machine Learning
- Computer Vision



Uncertainty Map

Alex Kendall, Yarin Gal: What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision? NIPS 2017: 5580-5590

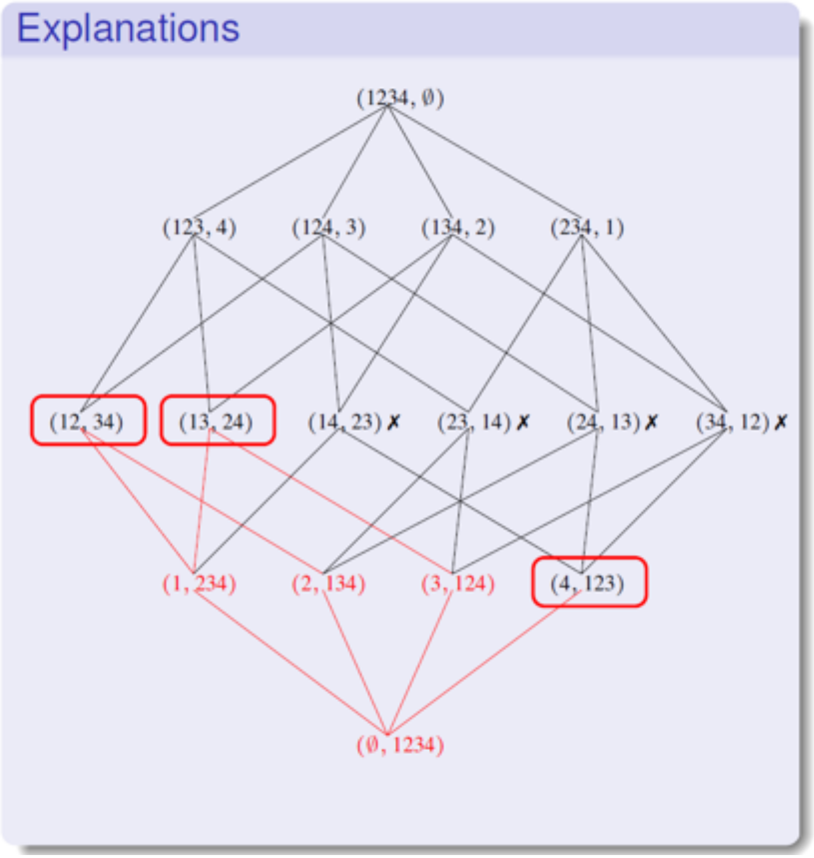


Saliency Map

Julius Adebayo, Justin Gilmer, Michael Muelly, Ian J. Goodfellow, Moritz Hardt, Been Kim: Sanity Checks for Saliency Maps. NeurIPS 2018: 9525-9536

Explanation in different AI fields

- Machine Learning
- Computer Vision
- Search and Constraint Satisfaction

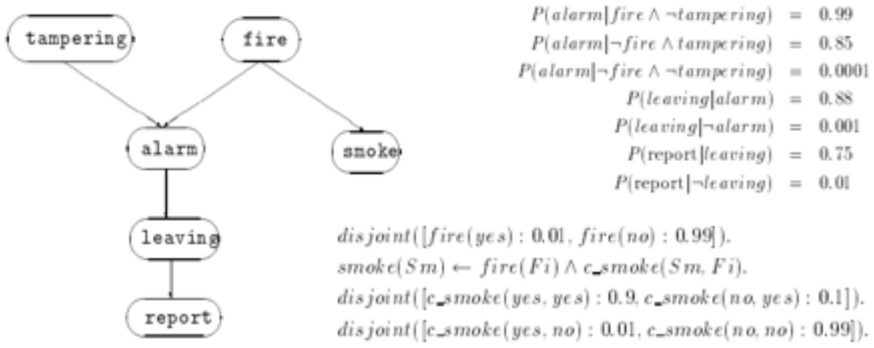


Constraints relaxation

Ulrich Junker: QUICKXPLAIN: Preferred Explanations and Relaxations for Over-Constrained Problems. AAAI 2004: 167-172

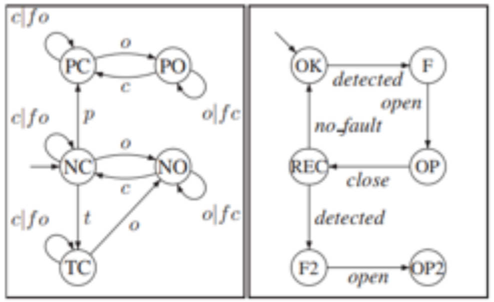
Explanation in different AI fields

- Machine Learning
- Computer Vision
- Knowledge Representation and Reasoning



Abduction Reasoning (in Bayesian Network)

David Poole: Probabilistic Horn Abduction and Bayesian Networks. Artif. Intell. 64(1): 81-129 (1993)

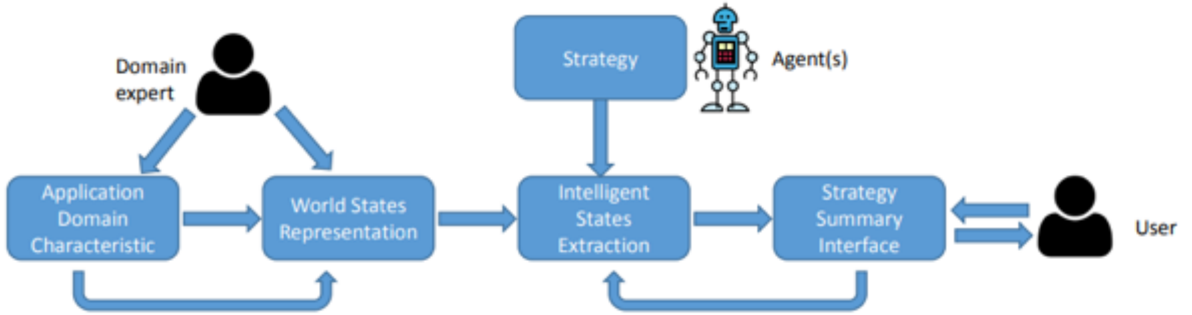


Diagnosis Inference

Alban Grastien, Patrik Haslum, Sylvie Thiébaux: Conflict-Based Diagnosis of Discrete Event Systems: Theory and Practice. KR 2012

Explanation in different AI fields

- Machine Learning
- Computer Vision
- Knowledge Representation and Reasoning
- Multi-agent Systems



Agent Strategy Summarization

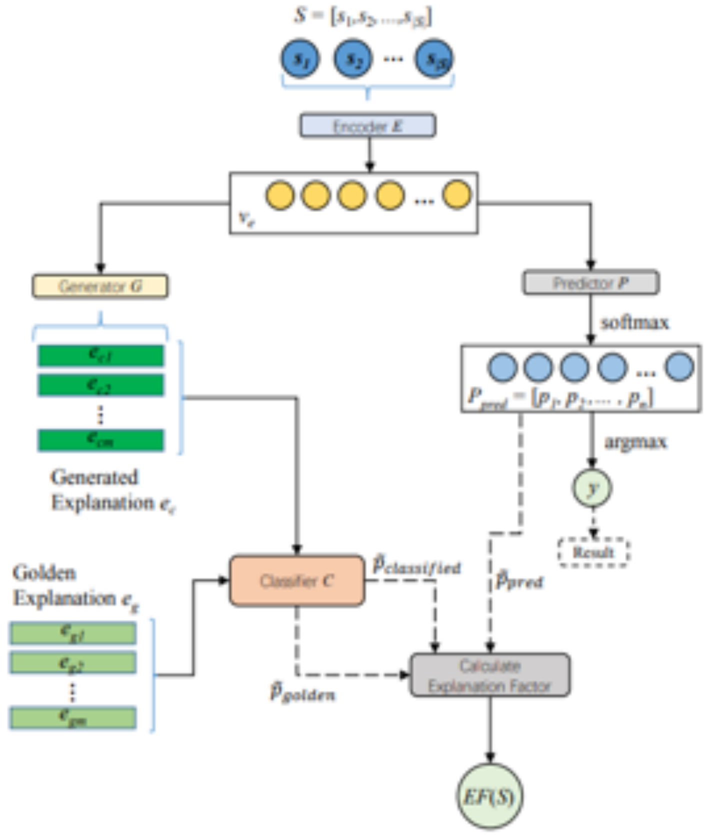
Ofra Amir, Finale Doshi-Velez, David Sarne: Agent Strategy Summarization. AAMAS 2018: 1203-1207

Explainable Agents

Joost Broekens, Maaike Harbers, Koen V. Hindriks, Karel van den Bosch, Catholijn M. Jonker, John-Jules Ch. Meyer: Do You Get It? User-Evaluated Explainable BDI Agents. MATES 2010: 28-39

Explanation in different AI fields

- Machine Learning
- Computer Vision
- Knowledge Representation and Reasoning
- Multi-agent Systems
- NLP

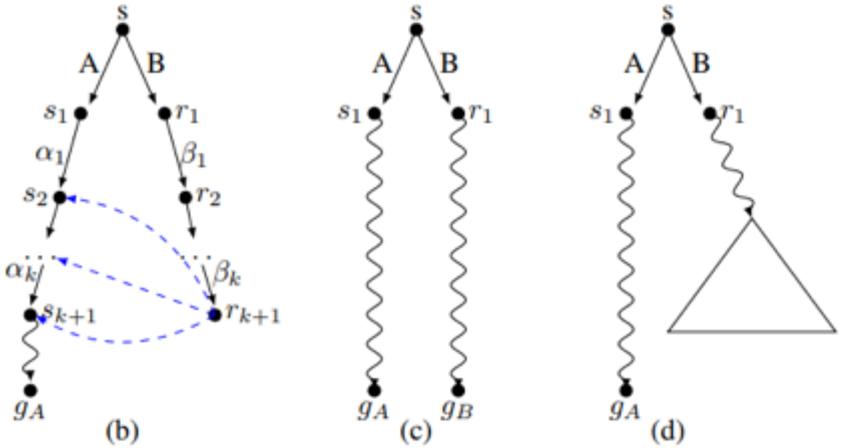


Explainable NLP

Hui Liu, Qingyu Yin, William Yang Wang: Towards Explainable NLP: A Generative Explanation Framework for Text Classification. CoRR abs/1811.00196 (2018)

Explanation in different AI fields

- Machine Learning
- Computer Vision
- Knowledge Representation and Reasoning
- Multi-agent Systems
- NLP
- Planning and Scheduling



Human-in-the-loop Planning

Maria Fox, Derek Long, Daniele Magazzeni: Explainable Planning. CoRR abs/1709.10256 (2017)

Explanation in different AI fields

- Machine Learning
- Computer Vision
- Knowledge Representation and Reasoning
- Multi-agent Systems
- NLP
- Planning and Scheduling
- Robotics

Robot: I have decided to turn left.

Human: Why did you do that?

Robot: I believe that the correct action is to turn left BECAUSE:
I'm being asked to go forward
AND This area in front of me was 20 cm higher than me *highlights area*
AND the area to the left has maximum protrusions of less than 5 cm *highlights area*
AND I'm tilted to the right by more than 5 degrees.
Here is a display of the path through the tree that lead to this decision. *displays tree*

Human: How confident are you in this decision?

Robot: The distribution of actions that reached this leaf node is shown in this histogram. *displays histogram* This action is predicted to be correct 67% of the time.

Human: Where did the threshold for the area in front come from?

Robot: Here is the histogram of all training examples that reached this leaf. 80% of examples where this area was above 20 cm predicted the appropriate action to be "drive forward".

From Decision Tree to human-friendly information

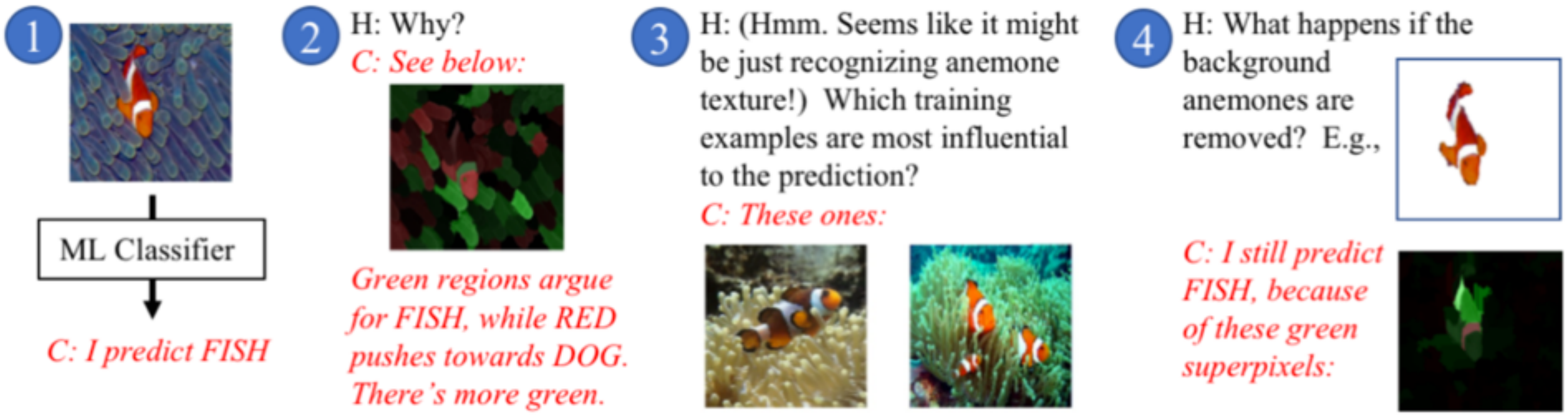
Raymond Ka-Man Sheh: "Why Did You Do That?" Explainable Intelligent Robots. AAAI Workshops 2017

Summarizing: the Need to Explain comes from ...

- User Acceptance & Trust [Lipton 2016, Ribeiro 2016, Weld and Bansal 2018]
- Legal
  - Conformance to ethical standards, fairness
  - *Right to be informed* [Goodman and Flaxman 2016, Wachter 2017]
  - Contestable decisions
- Explanatory Debugging [Kulesza et al. 2014, Weld and Bansal 2018]
  - Flawed performance metrics
  - Inadequate features
  - Distributional drift
- Increase Insightfulness [Lipton 2016]
  - Informativeness
  - Uncovering causality [Pearl 2009]

More ambitiously, explanation as *Machine-Human Conversation*

[Weld and Bansal 2018]



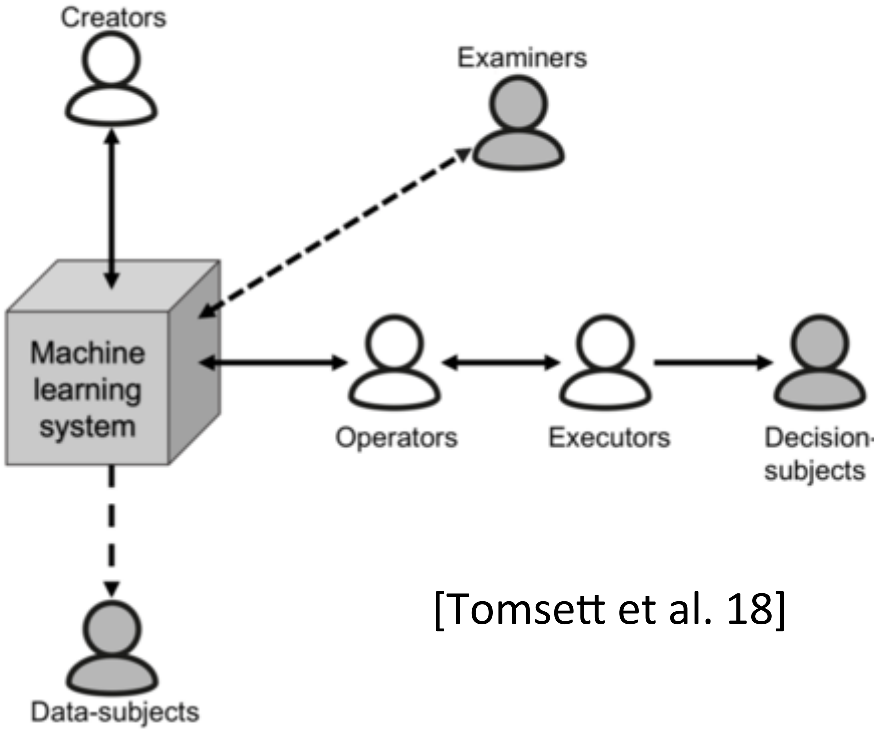
- Humans may have follow-up questions
- Explanations cannot answer all users' concerns

Role-based Interpretability

"~~Is the explanation interpretable~~?" → "*To whom* is the explanation interpretable?"

No Universally Interpretable Explanations!

- **End users** "Am I being treated fairly?"
  "Can I contest the decision?"
  "What could I do differently to get a positive outcome?"
- **Engineers, data scientists**: "Is my system working as designed?"
- **Regulators** " Is it compliant?"



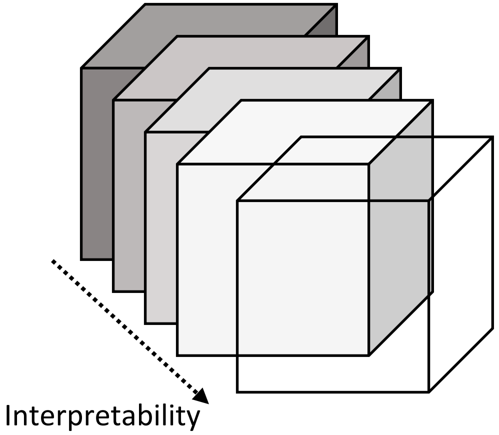
[Tomsett et al. 18]

An ideal explainer should model the *user background.*

[Tomsett et al. 2018, Weld and Bansal 2018, Poursabzi-Sangdeh 2018, Mittelstadt et al. 2019]

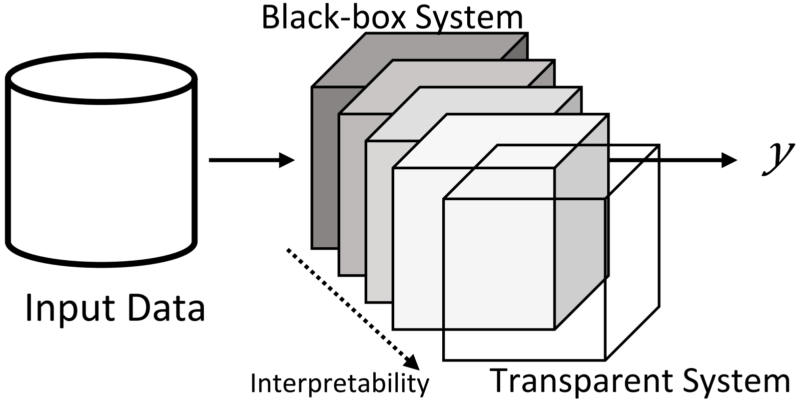
Evaluation: Interpretability as Latent Property

- Not directly measurable!
- Rely instead on *measurable outcomes*:
  - Any useful to individuals?
  - Can user estimate what a model will predict?
  - How much do humans follow predictions?
  - How well can people detect a mistake?
- No established benchmarks
- How to rank interpretable models? Different degrees of interpretability?

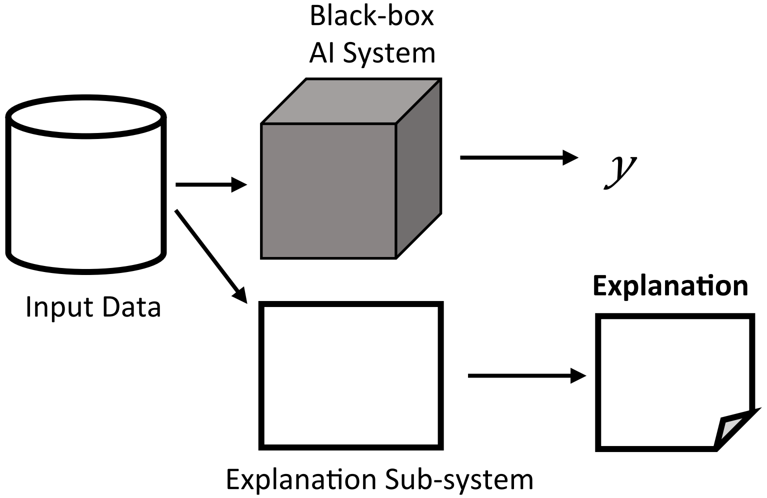


Explainable AI Systems

Transparent-by-design systems



Post-hoc Explanation (black-box explanation) systems



[Mittelstadt et al. 2018]

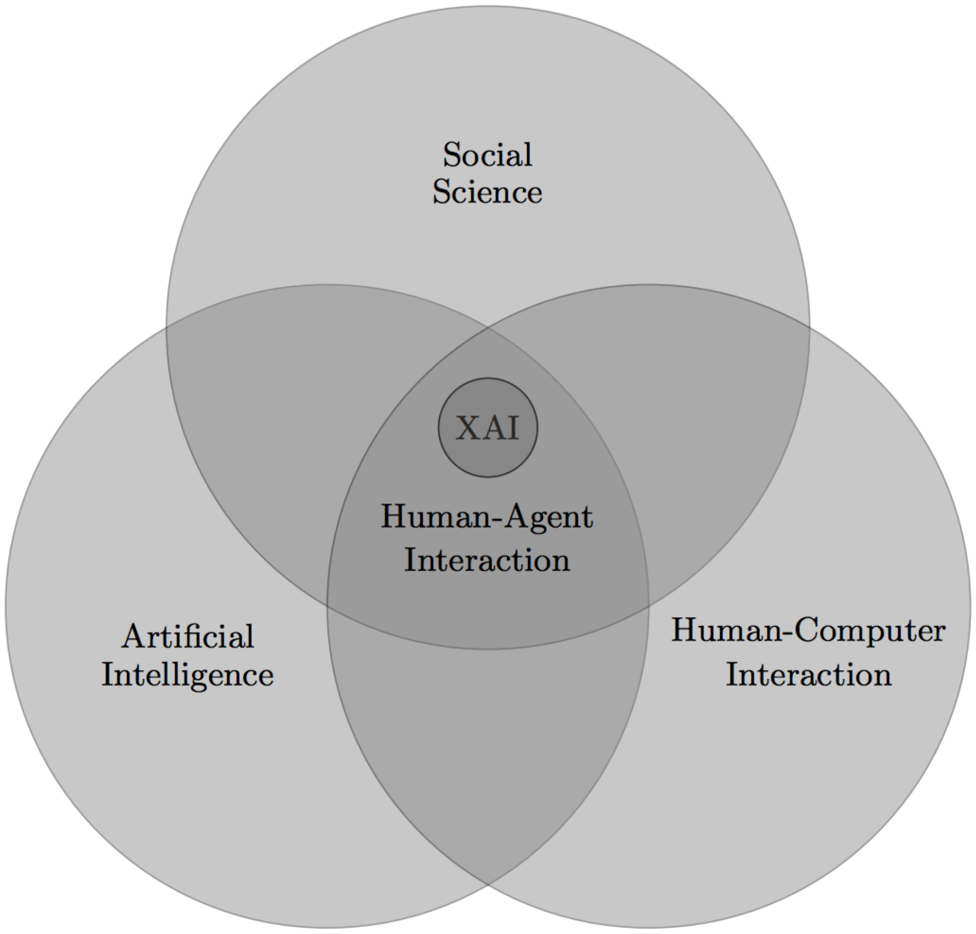
(Some) Desired Properties of Explainable AI Systems

- Informativeness
- Low cognitive load
- Usability
- Fidelity
- Robustness
- Non-misleading
- Interactivity /Conversational

[Lipton 2016, Doshi-velez and Kim 2017, Rudin 2018, Weld and Bansal 2018, Mittelstadt et al. 2019]

(thm) XAI is interdisciplinary

- For millennia, philosophers have asked the questions about what constitutes an explanation, what is the function of explanations, and what are their structure
- **[Tim Miller 2018]**



References

[Tim Miller 2018] Tim Miller Explanaition in Artificial Intelligence: Insight from Social Science

[Alvarez-Melis and Jaakkola 2018] Alvarez-Melis, David, and Tommi S. Jaakkola. "On the Robustness of Interpretability Methods." arXiv preprint arXiv:1806.08049 (2018).

[Chen and Rudin 2018]: Chaofan Chen and Cynthia Rudin. An optimization approach to learning falling rule lists. In Artificial Intelligence and Statistics (AISTATS), 2018.

[Doshi-Velez and Kim 2017] Doshi-Velez, Finale, and Been Kim. "Towards a rigorous science of interpretable machine learning." arXiv preprint arXiv:1702.08608 (2017).

[Goodman and Flaxman 2016] Goodman, Bryce, and Seth Flaxman. "European Union regulations on algorithmic decision-making and a" right to explanation"." arXiv preprint arXiv: 1606.08813 (2016).

[Freitas 2014] Freitas, Alex A. "Comprehensible classification models: a position paper." ACM SIGKDD explorations newsletter 15.1 (2014): 1-10.

[Goodman and Flaxman 2016] Goodman, Bryce, and Seth Flaxman. "European Union regulations on algorithmic decision-making and a" right to explanation"." arXiv preprint arXiv: 1606.08813 (2016).

[Gunning 2017] Gunning, David. "Explainable artificial intelligence (xai)." Defense Advanced Research Projects Agency (DARPA), nd Web (2017).

[Hind et al. 2018] Hind, Michael, et al. "Increasing Trust in AI Services through Supplier's Declarations of Conformity." arXiv preprint arXiv:1808.07261 (2018).

[Kulesza et al. 2014] Kulesza, Todd, et al. "Principles of explanatory debugging to personalize interactive machine learning." Proceedings of the 20th international conference on intelligent user interfaces. ACM, 2015.

[Lipton 2016] Lipton, Zachary C. "The mythos of model interpretability. Int. Conf." Machine Learning: Workshop on Human Interpretability in Machine Learning. 2016.

[Mittelstatd et al. 2019] Mittelstadt, Brent, Chris Russell, and Sandra Wachter. "Explaining explanations in AI." arXiv preprint arXiv:1811.01439 (2018).

[Poursabzi-Sangdeh 2018] Poursabzi-Sangdeh, Forough, et al. "Manipulating and measuring model interpretability." arXiv preprint arXiv:1802.07810 (2018).

[Rudin 2018] Rudin, Cynthia. "Please Stop Explaining Black Box Models for High Stakes Decisions." arXiv preprint arXiv:1811.10154 (2018).

[Wachter et al. 2017] Wachter, Sandra, Brent Mittelstadt, and Luciano Floridi. "Why a right to explanation of automated decision-making does not exist in the general data protection regulation." International Data Privacy Law 7.2 (2017): 76-99.

[Weld and Bansal 2018] Weld, D., and Gagan Bansal. "The challenge of crafting intelligible intelligence." Communications of ACM (2018).

[Yin 2012] Lou, Yin, Rich Caruana, and Johannes Gehrke. "Intelligible models for classification and regression." Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, (2012).

Explainable Machine Learning

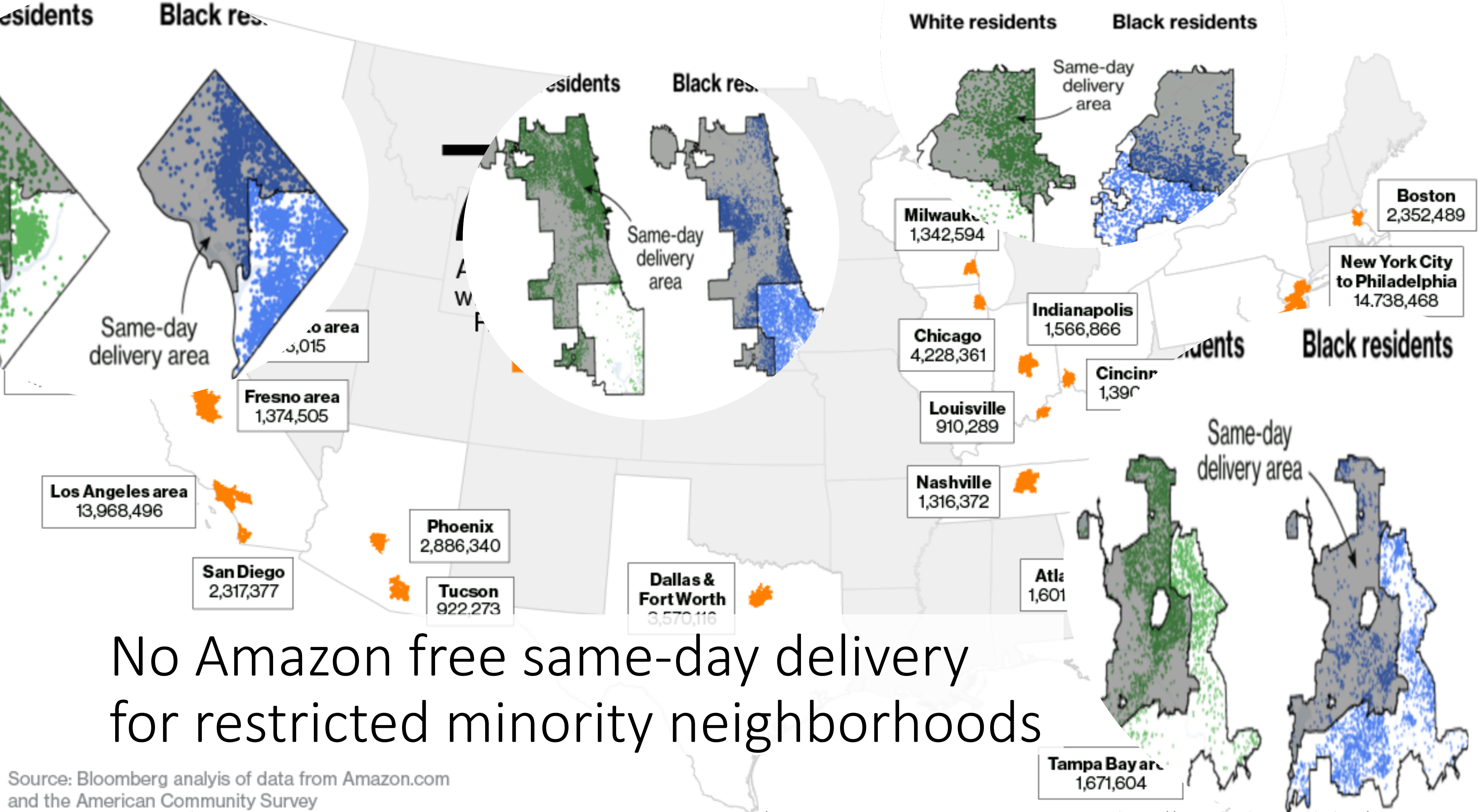
What is a Black Box Model?

A ***black box*** is a model, whose internals are either unknown to the observer or they are known but uninterpretable by humans.

- 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.

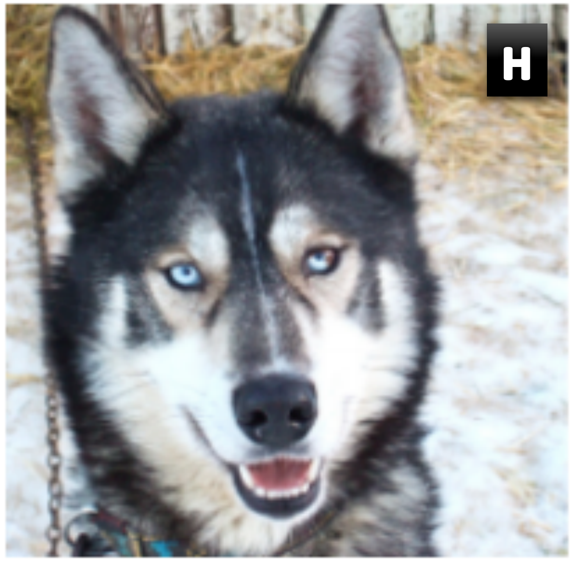
Bias in Machine Learning

COMPAS recidivism black bias

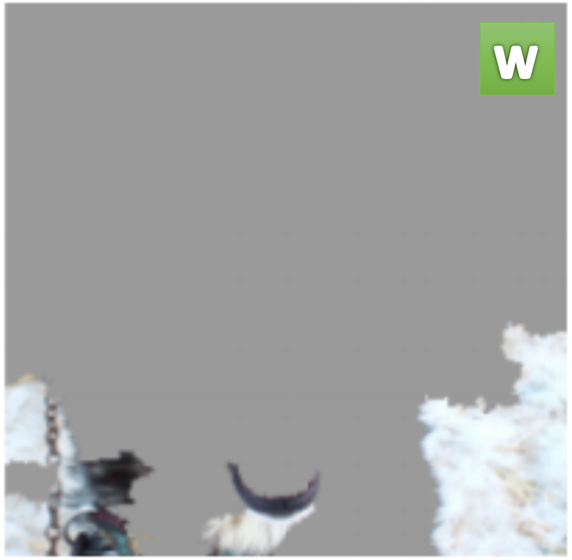


No Amazon free same-day delivery for restricted minority neighborhoods

The background bias



(a)Husky classified as wolf



(b)Explanation

Ribeiro et al., KDD 2016

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

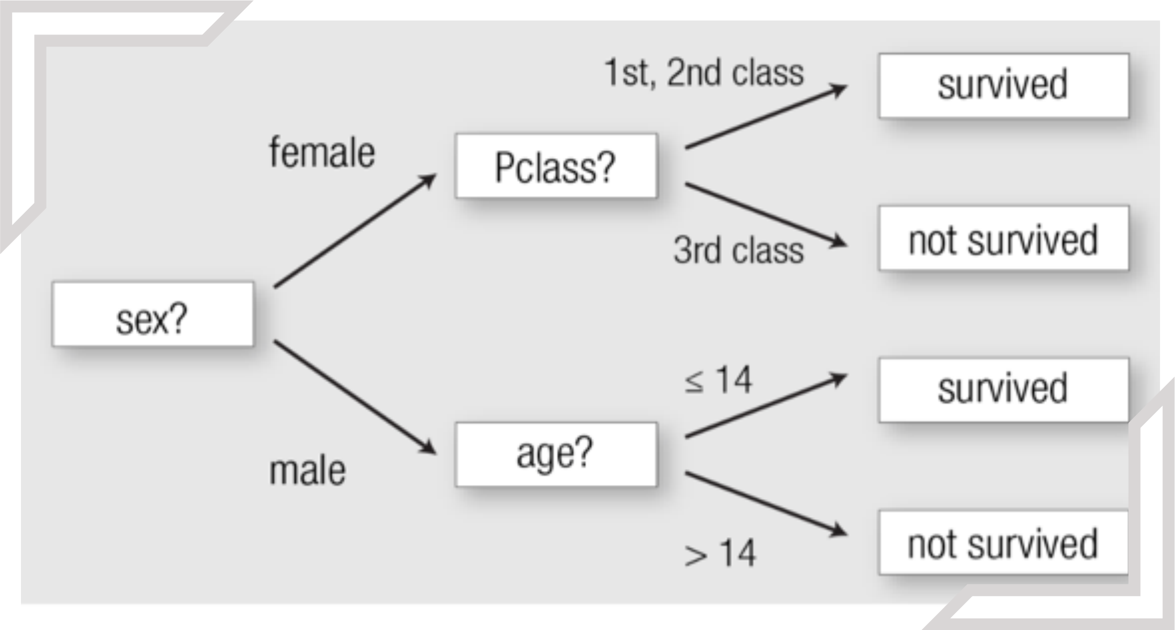
- ***Fairness***: the model guarantees the protection of groups against discrimination.
- ***Privacy***: the model does not reveal sensitive information about people.
- ***Respect Monotonicity***: the increase of the values of an attribute either increase or decrease in a monotonic way the probability 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.

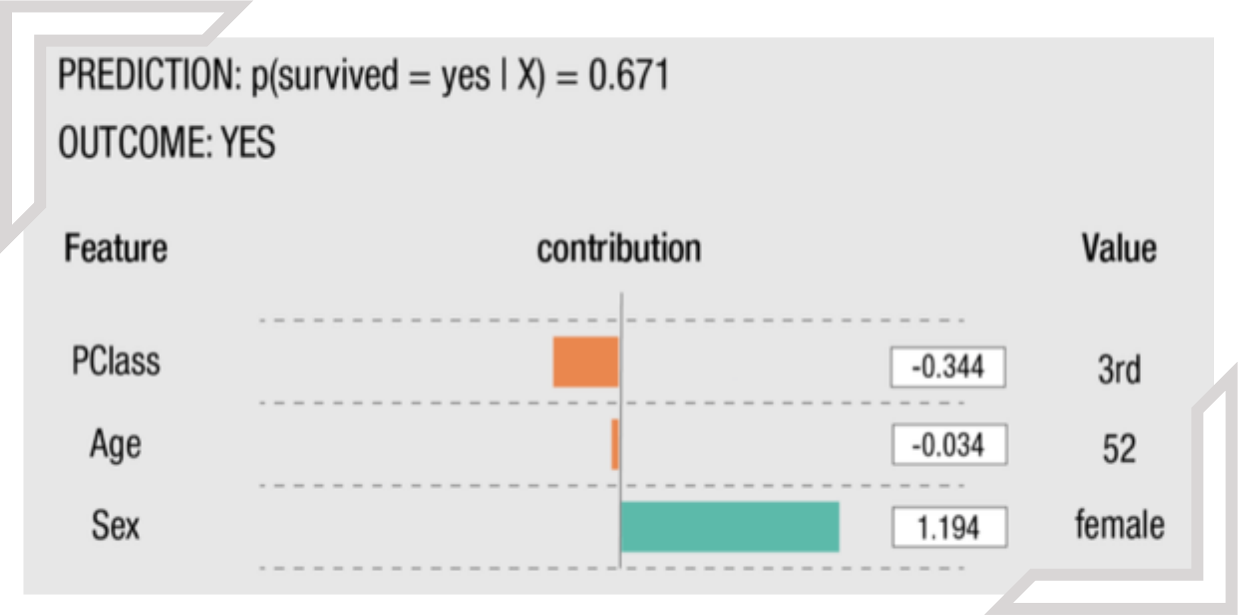
Desiderata of an Interpretable Model

- ***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

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

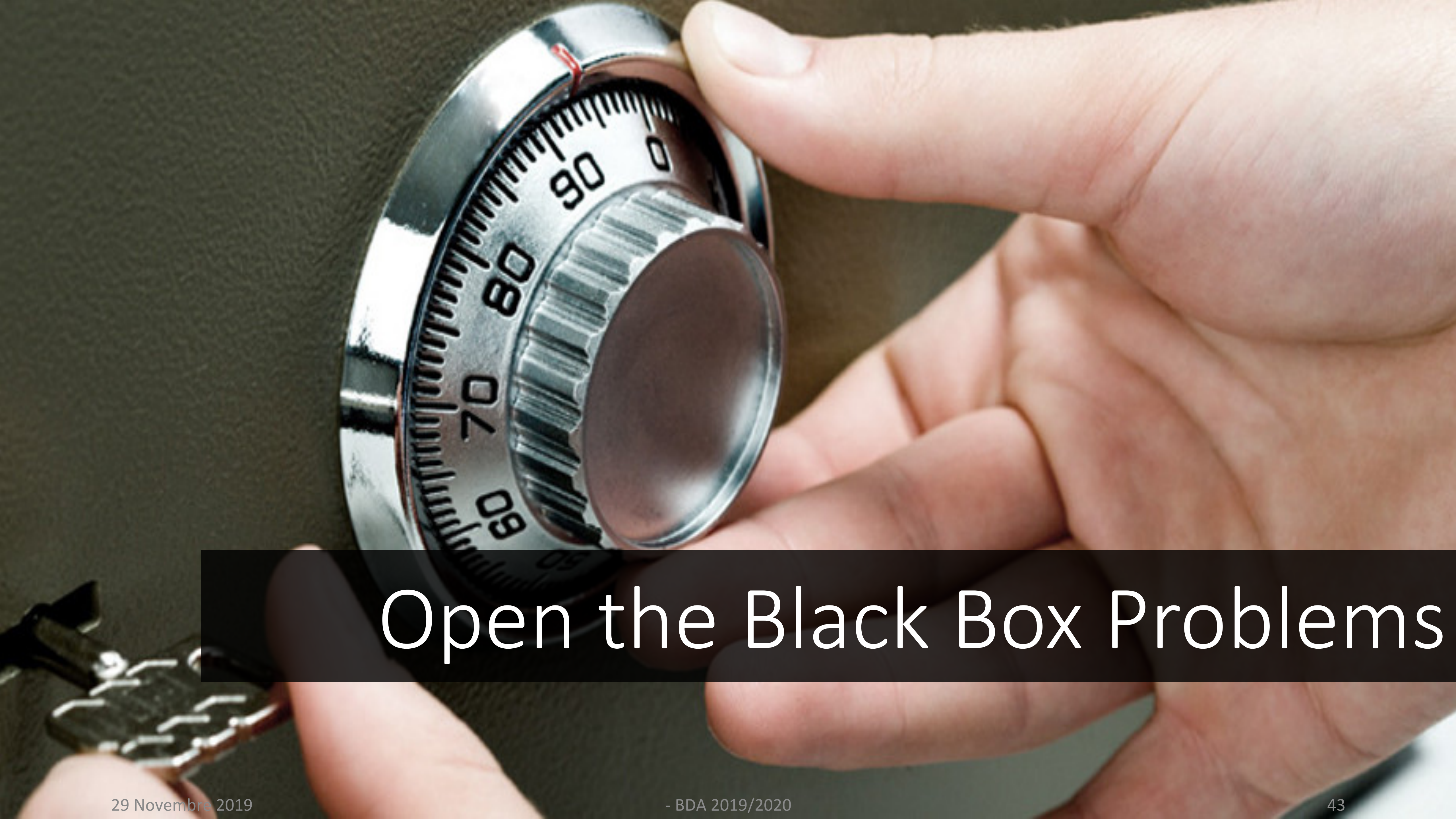
Rules

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.

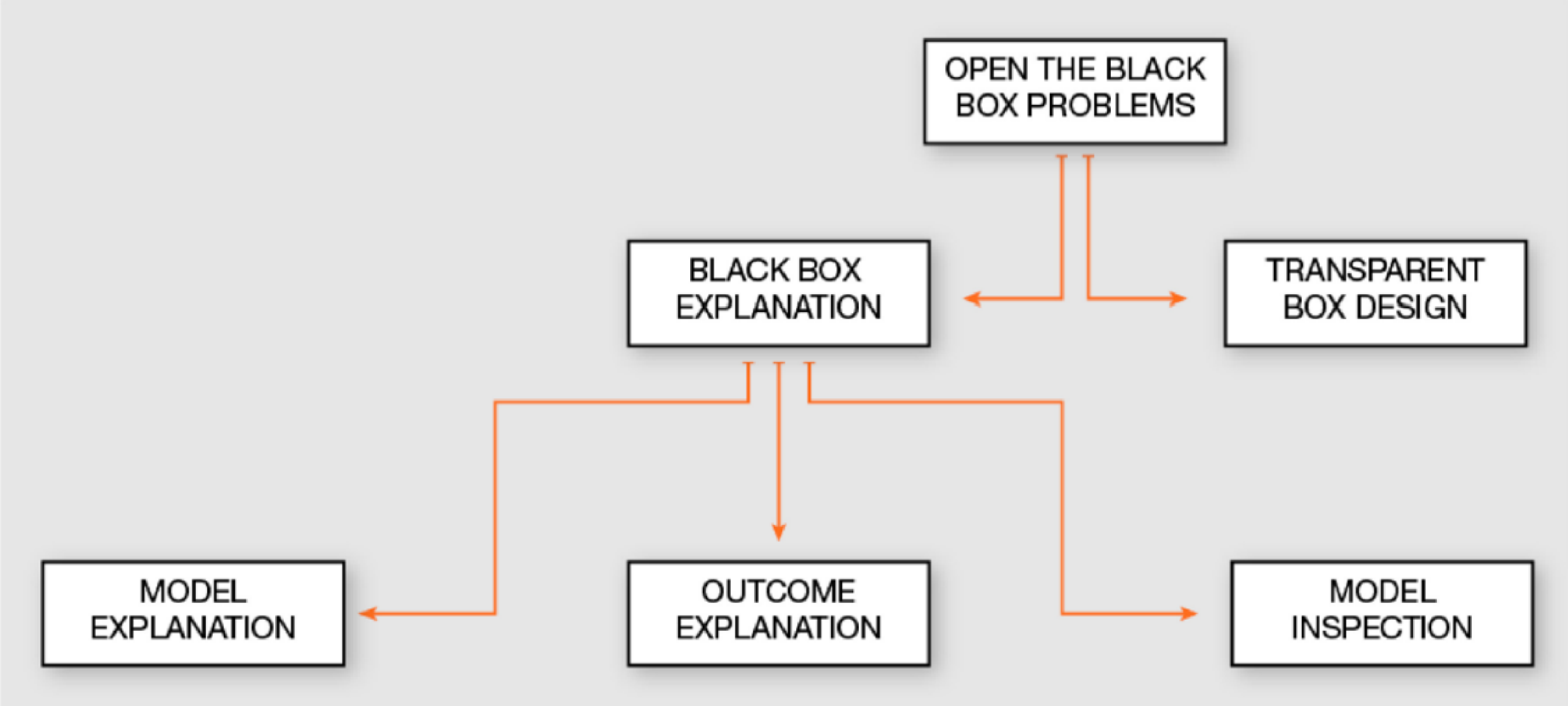
- Linear Model: number of non zero weights in the model.
- Rule: number of attribute-value pairs in condition.
- Decision Tree: estimating the complexity of a tree can be hard.

- 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.

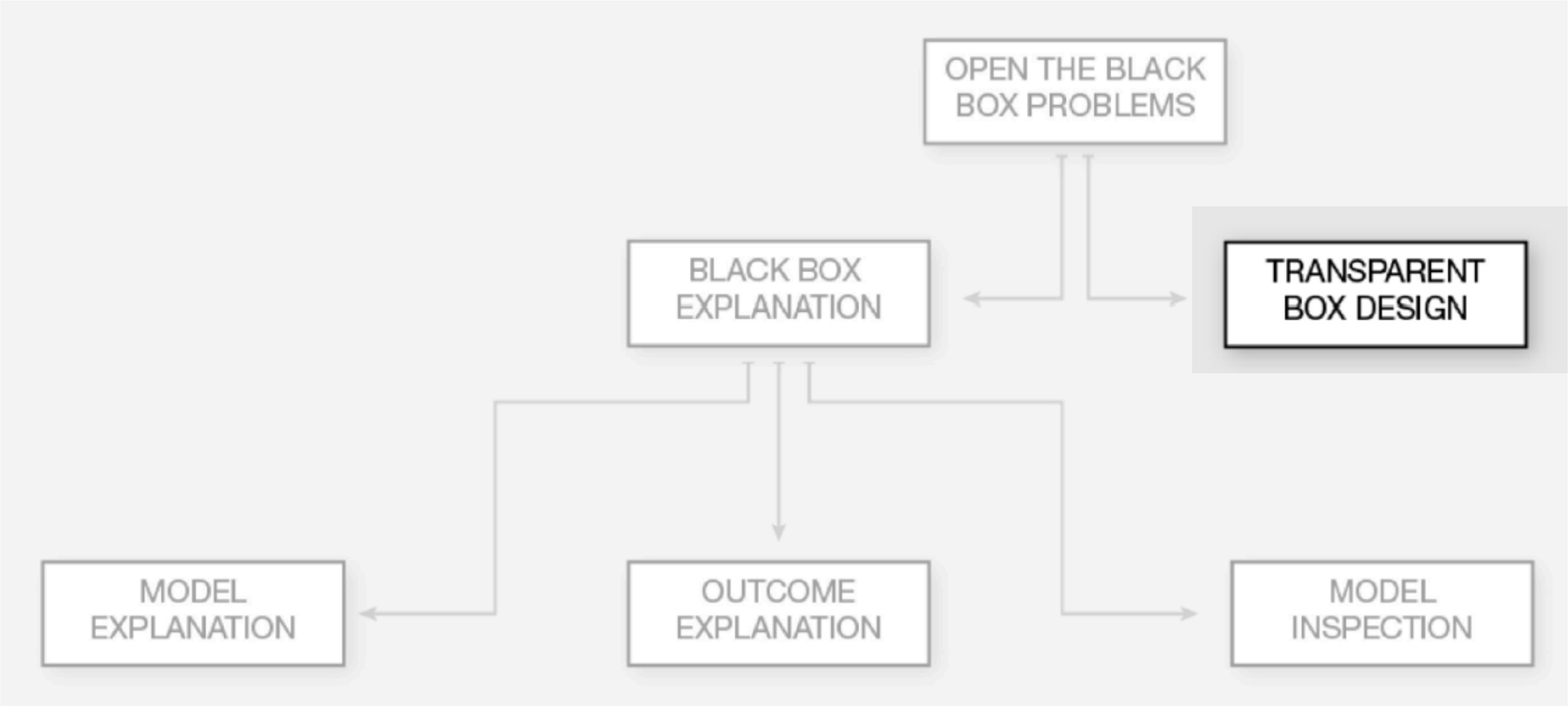


Open the Black Box Problems

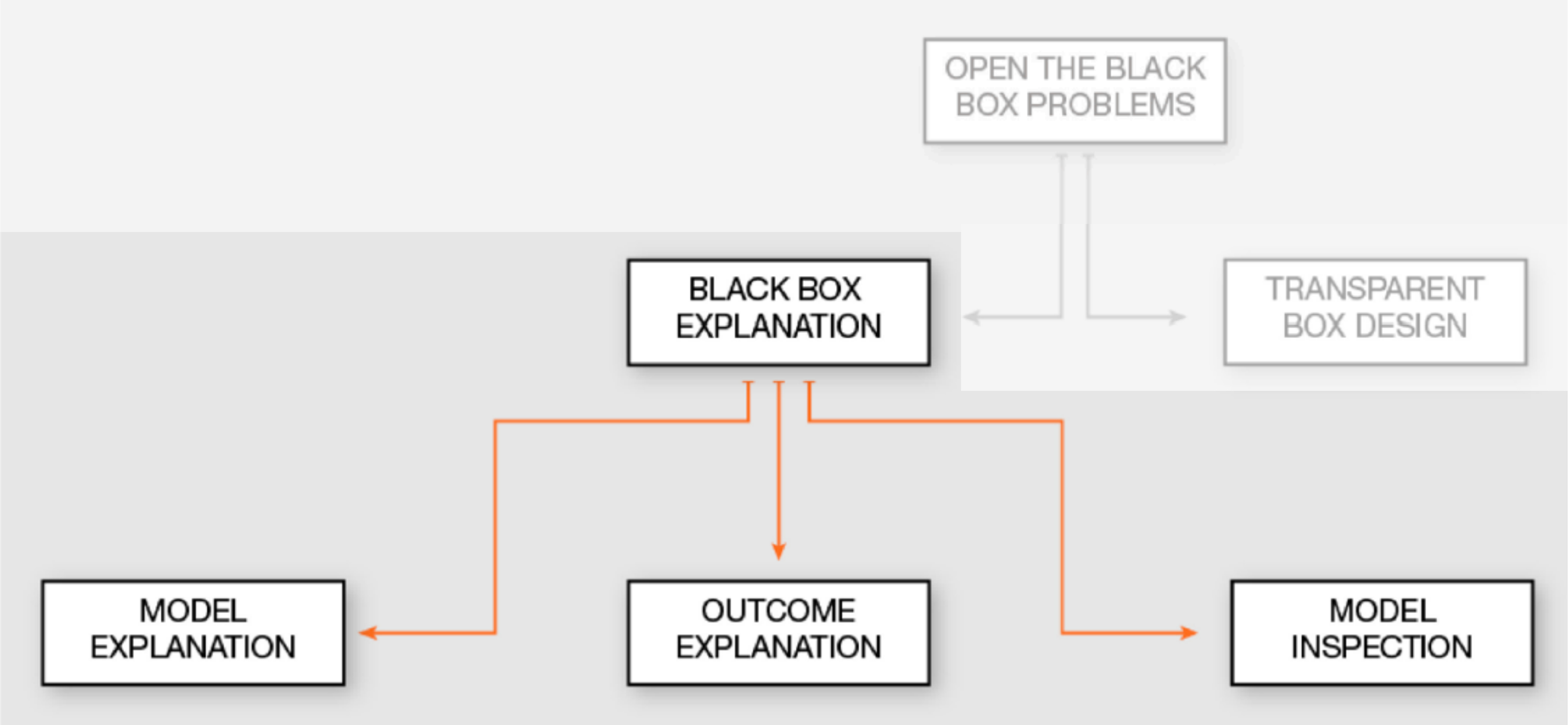
Problems Taxonomy



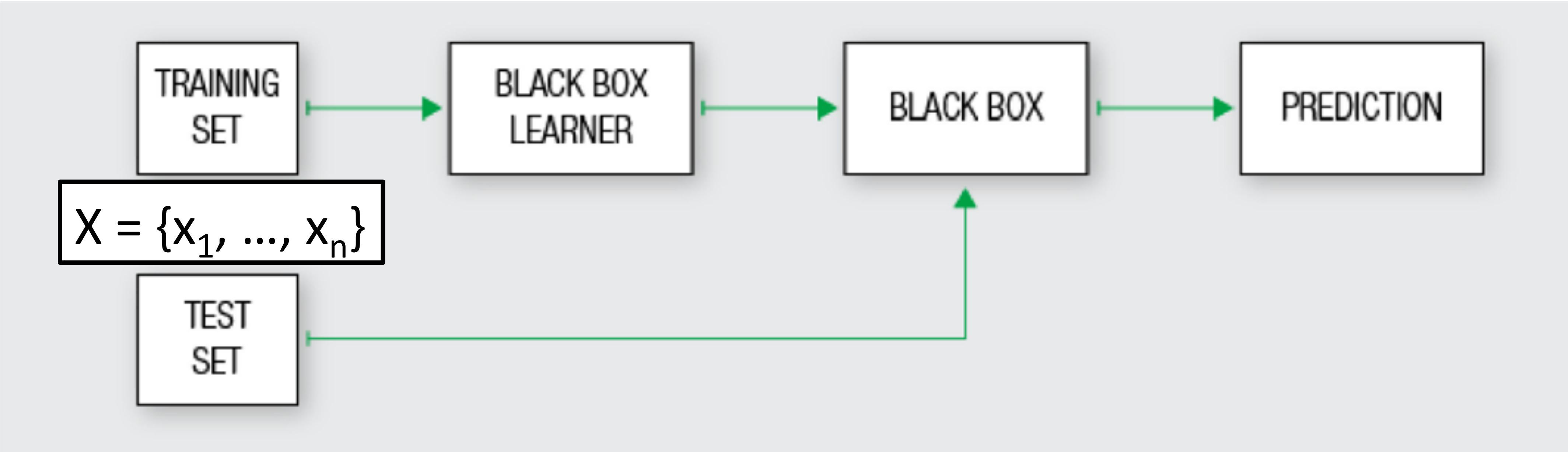
XbD – eXplanation by Design



BBX - Black Box eXplanation



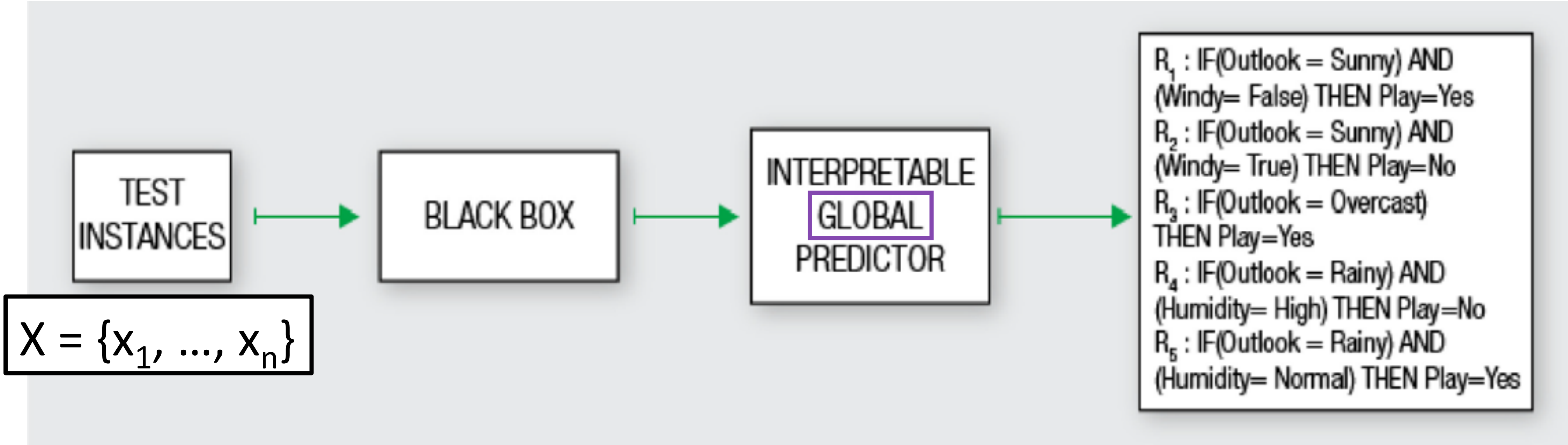
ML Problem



$X = \{x_1, \ldots, x_n\}$

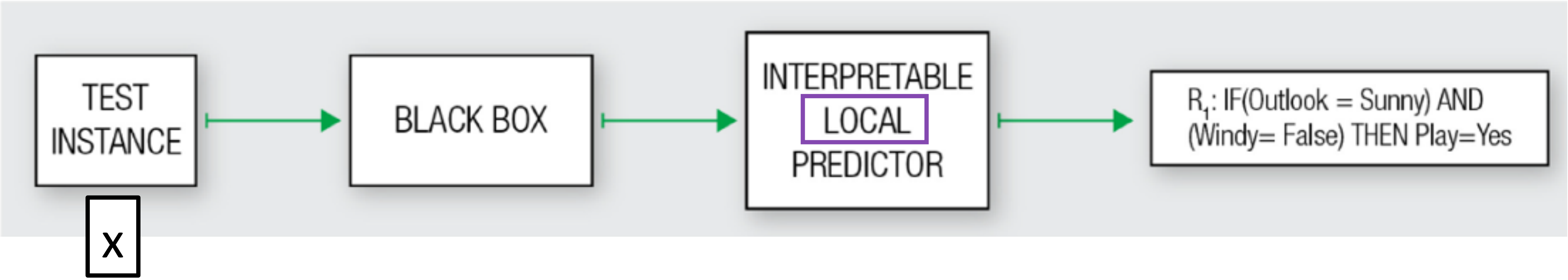
Model Explanation Problem

Provide an interpretable model able to mimic the ***overall logic/behavior*** of the black box and to explain its logic.



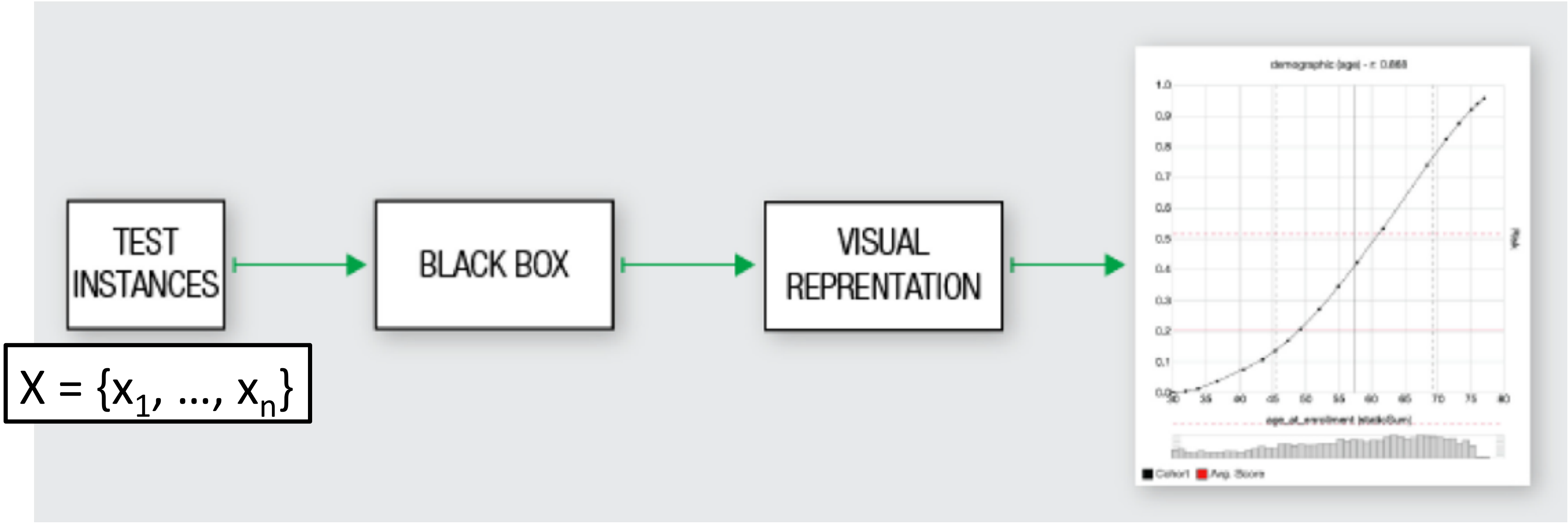
Outcome Explanation Problem

Provide an interpretable outcome, i.e., an ***explanation*** for the outcome of the black box for a ***single instance***.



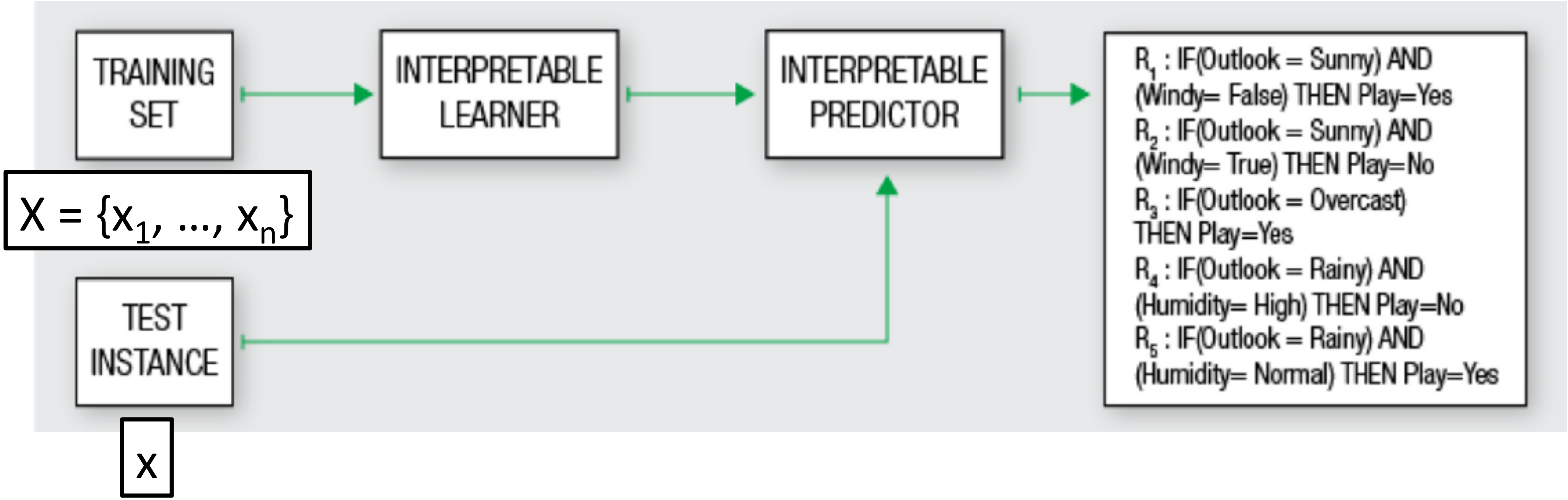
Model Inspection Problem

Provide a representation (visual or textual) for understanding either how the black box model works or why the black box returns certain predictions more likely than others.

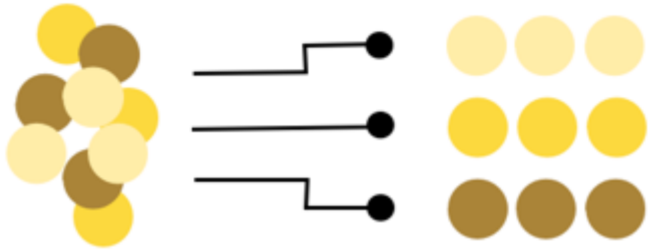


Transparent Box Design Problem

Provide a model which is locally or globally interpretable on its own.

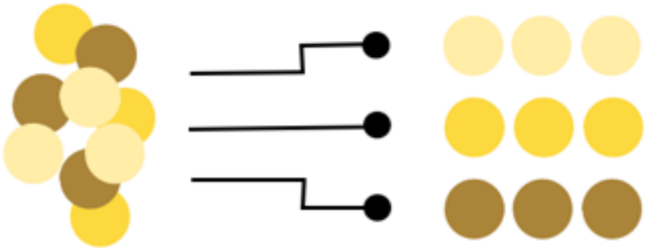


Categorization



- The type of ***problem***
- The type of ***black box model*** that the explanator is able to open
- The type of ***data*** used as input by the black box model
- The type of ***explanator*** adopted to open the black box

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)

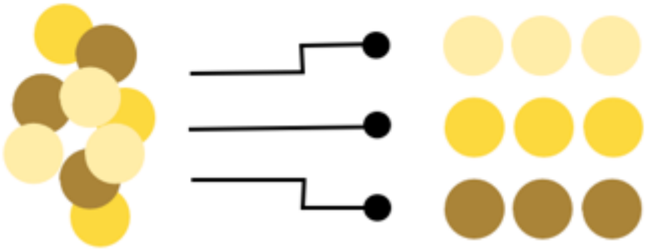
2000 rows all told

Tabular
(TAB)

Images
(IMG)

Text
(TXT)

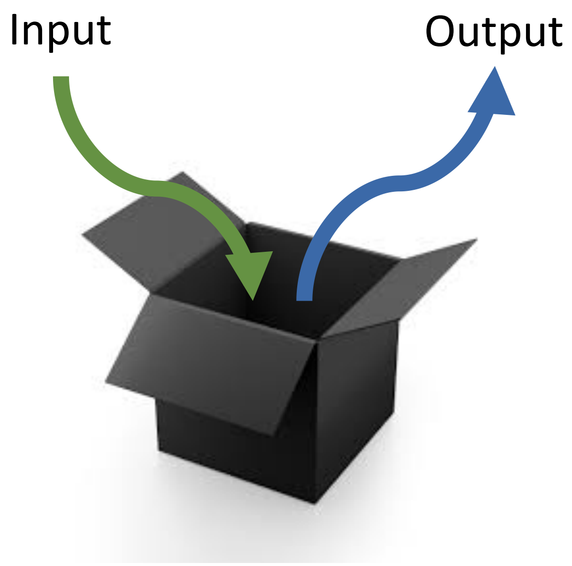
Explanators



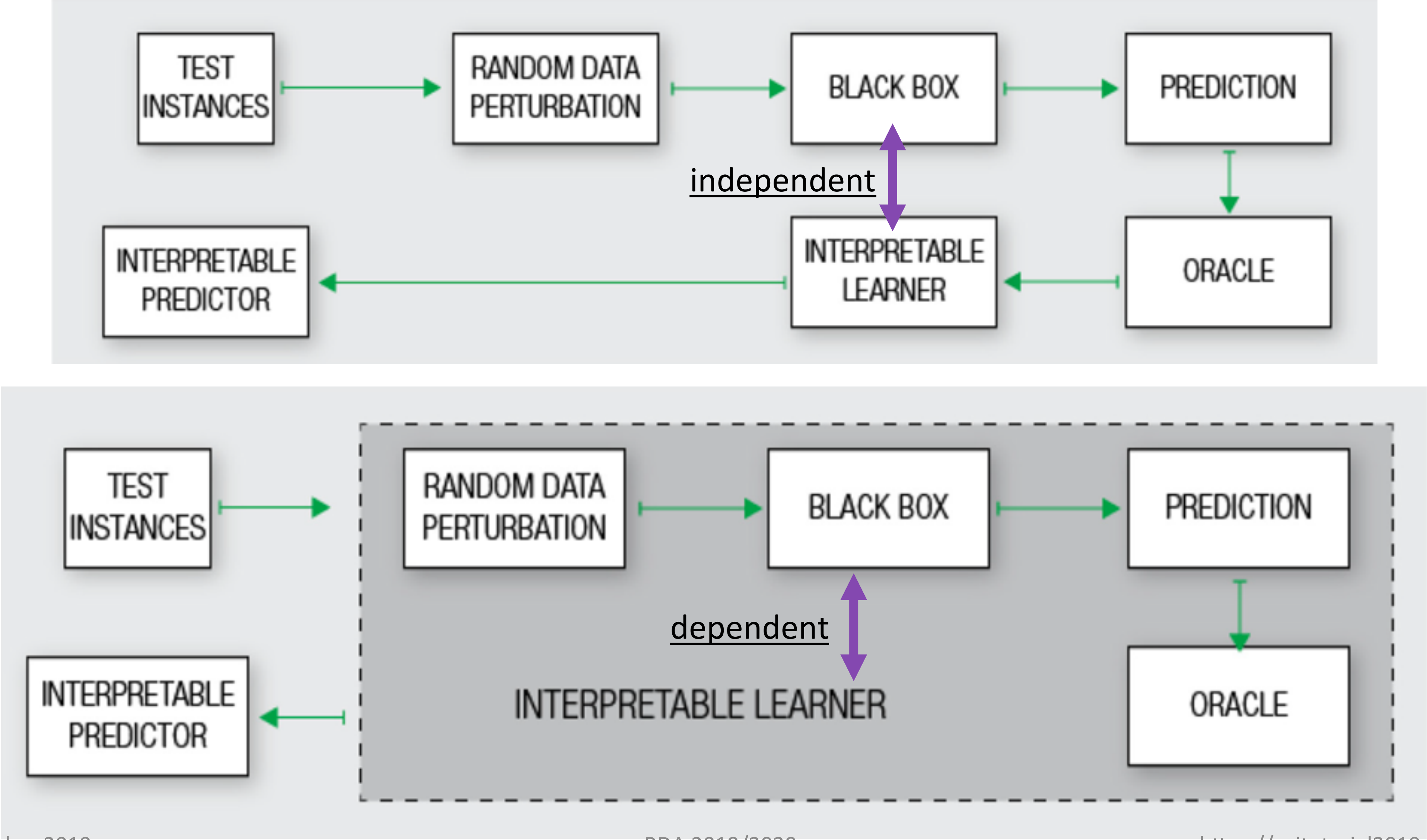
- Decision Tree (***DT***)
- Decision Rules (***DR***)
- Features Importance (***FI***)
- Saliency Mask (***SM***)
- Sensitivity Analysis (***SA***)
- Partial Dependence Plot (***PDP***)
- Prototype Selection (***PS***)
- Activation Maximization (***AM***)

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



Model-Agnostic vs Model-Specific



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		✓	✓		✓

Solving The Model Explanation Problem

Global Model Explainers

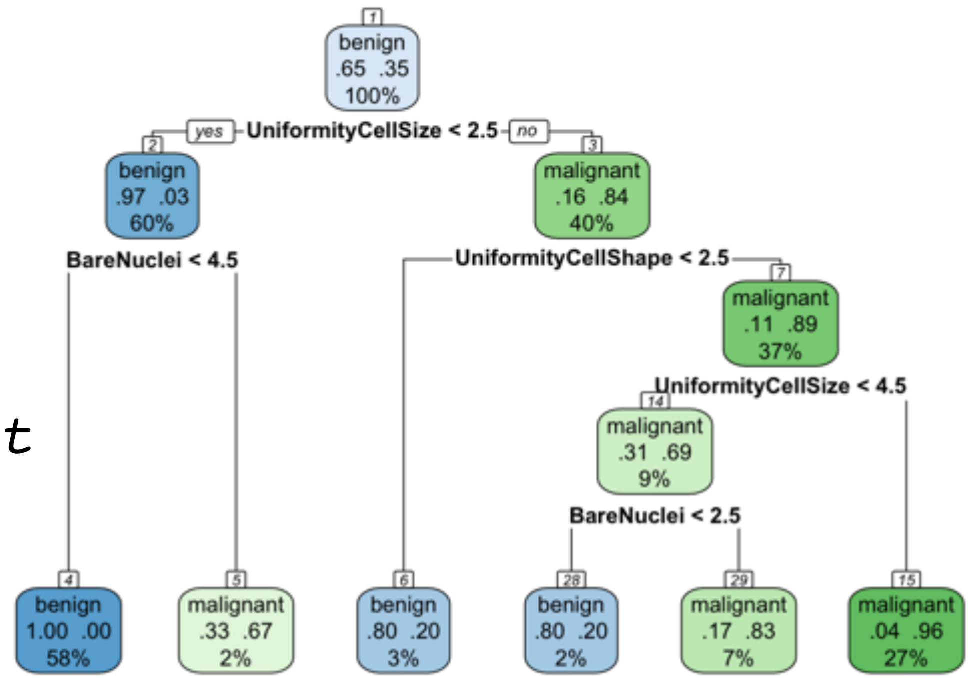
- Explanator: DT
    - Black Box: NN, TE
    - Data Type: TAB
- Explanator: DR
    - Black Box: NN, SVM, TE
    - Data Type: TAB
- Explanator: FI
    - Black Box: AGN
    - Data Type: TAB

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

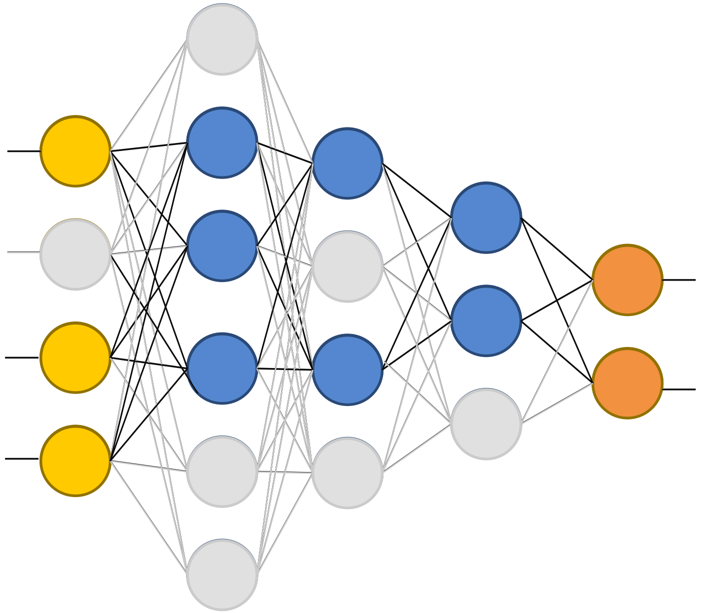
```
T = root_of_the_tree()
Q = <T, X, {}>
while Q not empty & size(T) < limit
      N, X_N, C_N = pop(Q)
      Z_N = random(X_N, C_N)
      y_Z = b(Z), y = b(X_N)
      if same_class(y ∪ y_Z)
            continue
      S = best_split(X_N ∪ Z_N, y ∪ y_Z)
      S'= best_m-of-n_split(S)
      N = update_with_split(N, 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>)
```

black box auditing



- 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.

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.