Explainable AI:

From Theory to Motivation, Applications and Challenges

Lecturers: **Fosca Giannotti (ISTI-CNR), Dino Pedreschi (University of Pisa)**

Contributors: S. Rinzivillo, R. Guidotti (ISTI-CNR)

A. Monreale, F. Turini, S. Ruggieri (University of Pisa)

http://ai4eu.org/

http://www.sobigdata.eu/

http://www.humane-ai.eu/

ERC-AdG-2019 "Science & technology for the eXplanation of AI decision making"


Vous préférez
un conseiller
qui répond
humainement

ou une machine
qui répond
machinalement ?

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

[Floridi 2019

Material based on (our) XAI Tutorial at AAAI2019

https://xaitutorial2019.github.io/

Disclaimer:

- **As MANY interpretations as research areas** (check out work in Machine Learning vs Reasoning community)
- Not an exhaustive survey! Focus is on some promising approaches
- Massive body of literature (growing in time)
- Multi-disciplinary (AI – all areas, HCI, social sciences)
- Many domain-specific works hard to uncover
- Many papers do not include the keywords explainability/interpretability!

Motivating Example (1)

- Criminal Justice
  - People wrongly denied
  - Recidivism prediction
  - Unfair Police dispatch

Opinion

The New York Times

OP-ED CONTRIBUTOR

When a Computer Program Keeps You in Jail

By Rebecca Wexler

June 13, 2017

nytimes.com/2017/06/13/opinion/how-computers-are-harming-criminal-justice.html

How We Analyzed the COMPAS Recidivism Algorithm

by Jeff Larson, Surya Mattu, Lauren Kirchner and Julia Angwin

May 23, 2016

propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm

ACLU GET UPDATES /

STATEMENT OF CONCERN ABOUT PREDIC POLICING BY ACLU AND 16 CIVIL RIGHTS PRIVACY, RACIAL JUSTICE, AND TECHNOLOGY ORGANIZATIONS

aclu.org/other/statement-concern-about-predictive-policing-aclu-and-16-civil-rights-privacy-racial-justice

[Rudin 2018]

Motivating Example (2)

- Finance:
  - Credit scoring, loan approval
  - Insurance quotes

The Big Read Artificial intelligence + Add to myFT

Insurance: Robots learn the business of covering risk

Artificial intelligence could revolutionise the industry but may also allow clients to calculate if they need protection

Save

Oliver Ralph MAY 16, 2017 24

https://www.ft.com/content/e07cee0c-3949-11e7-821a-6027b8a20f23

community.fico.com/s/explainable-machine-learning-challenge

Motivating Example (3)

- 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.

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

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.

Patricia Hannon ,https://med.stanford.edu/news/all-news/2018/03/researchers-say-use-of-ai-in-medicine-raises-ethical-questions.html

Intelligible Models for HealthCare: Predicting Pneumonia Risk and Hospital 30-day Readmission

Rich Caruana
Microsoft Research
rcaruana@microsoft.com

Yin Lou
LinkedIn Corporation
ylou@linkedin.com

Johannes Gehrke
Microsoft
johannes@microsoft.com

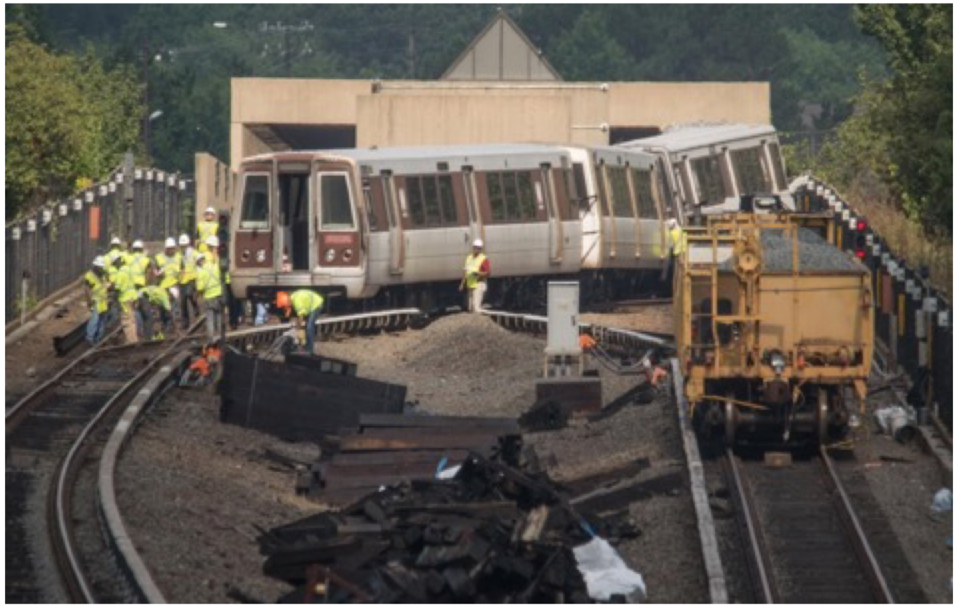
Paul Koch
Microsoft Research
paulkoch@microsoft.com

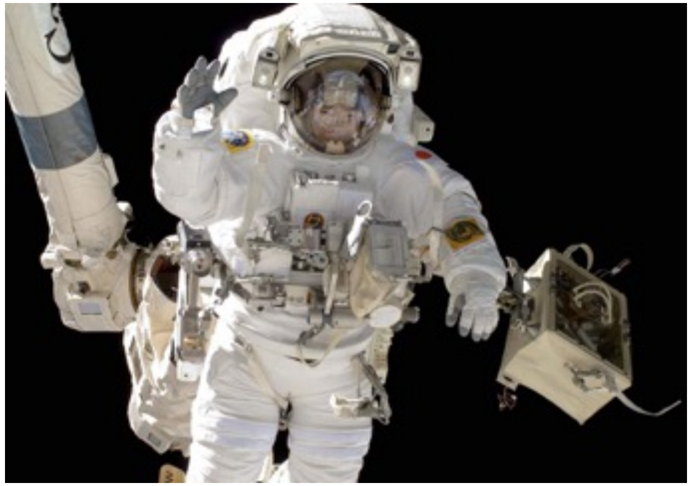
Marc Sturm
NewYork-Presbyterian Hospital
mas9161@nyp.org

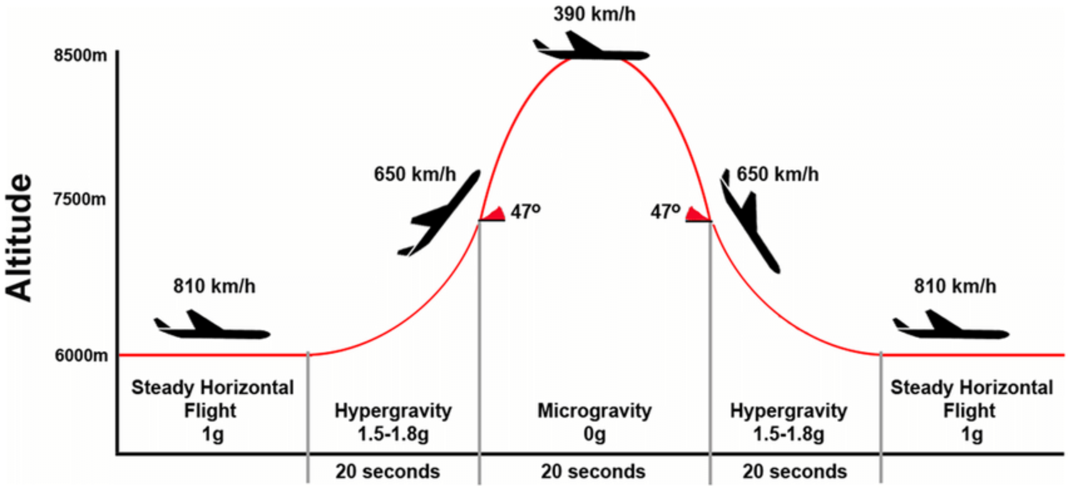
Noémie Elhadad
Columbia University
noemie.elhadad@columbia.edu

Motivation (4)

- Critical Systems

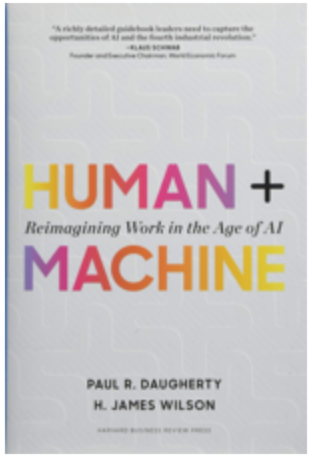




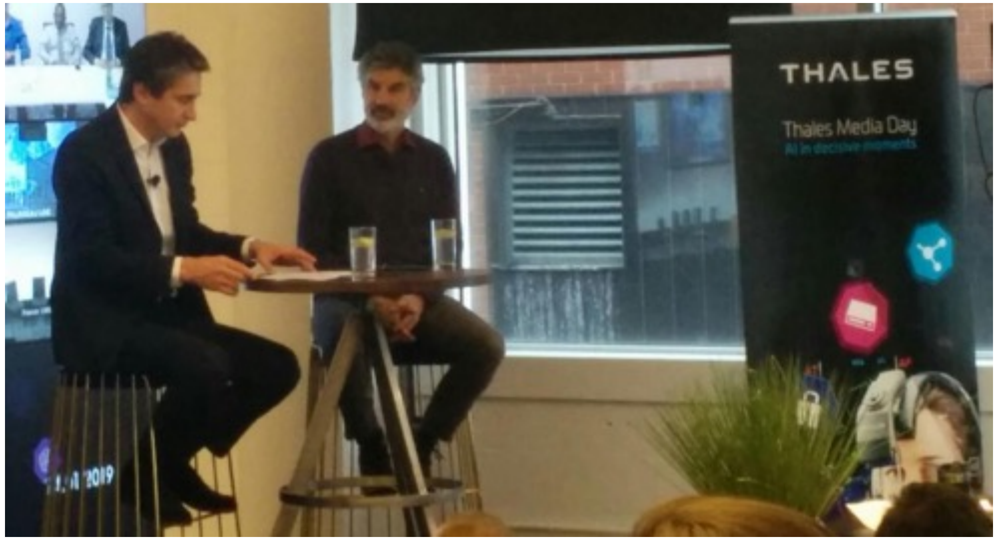


[Caruana et al. 2015, Holzinger et al. 2017, Magnus et al. 2018]

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.

Tutorial Outline (1)

- **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
- **Applications**

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.

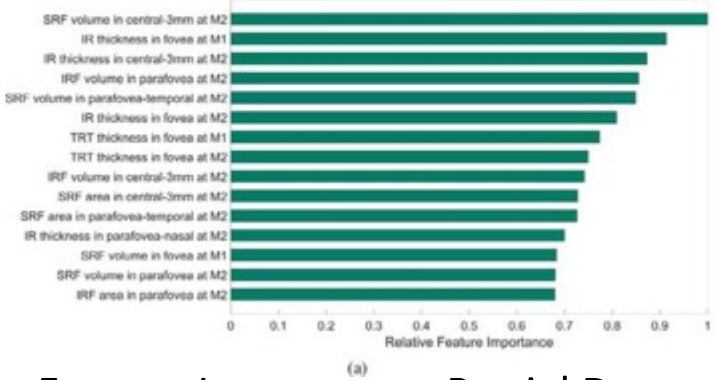
[Lepri et al. 2018] Lepri, Bruno, et al. "Fair, Transparent, and Accountable Algorithmic Decision-making Processes." Philosophy & Technology (2017): 1-17.

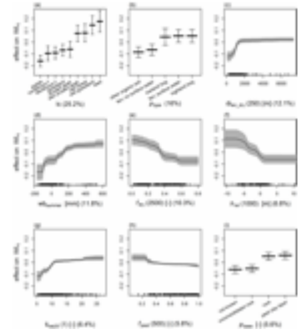
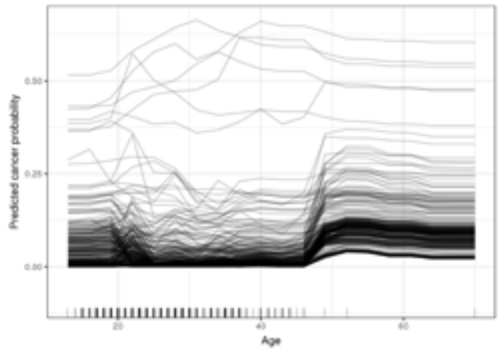
[Floridi et al**. 2019] Floridi,** Luciano and Josh Cowls "A Unified Framework of Five Principles for AI in Society". Harvard Data Science Review, 1, 2019

Explanation in AI

Overview of explanation in different AI fields (1)

- Machine Learning

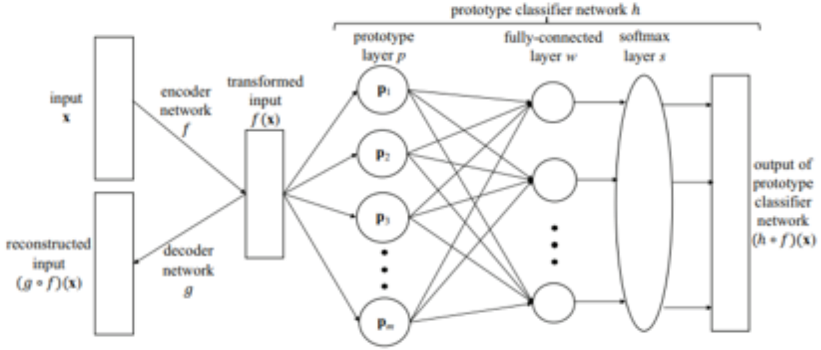


Feature Importance, Partial Dependence Plot, Individual Conditional Expectation

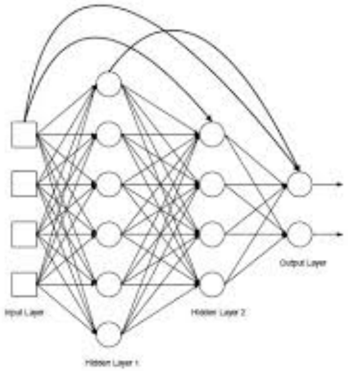
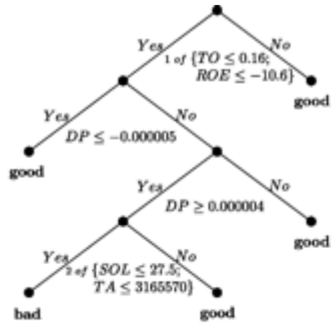
Interpretable Models:

- Linear regression,
- Logistic regression,
- Decision Tree,
- Naive Bayes,
- KNNs



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

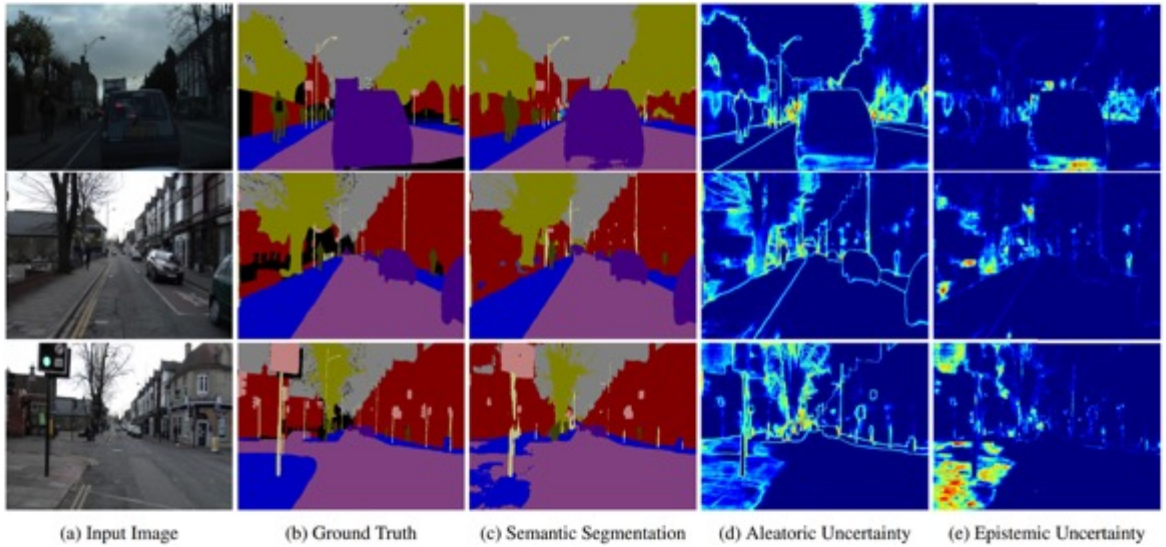



Surogate Model

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

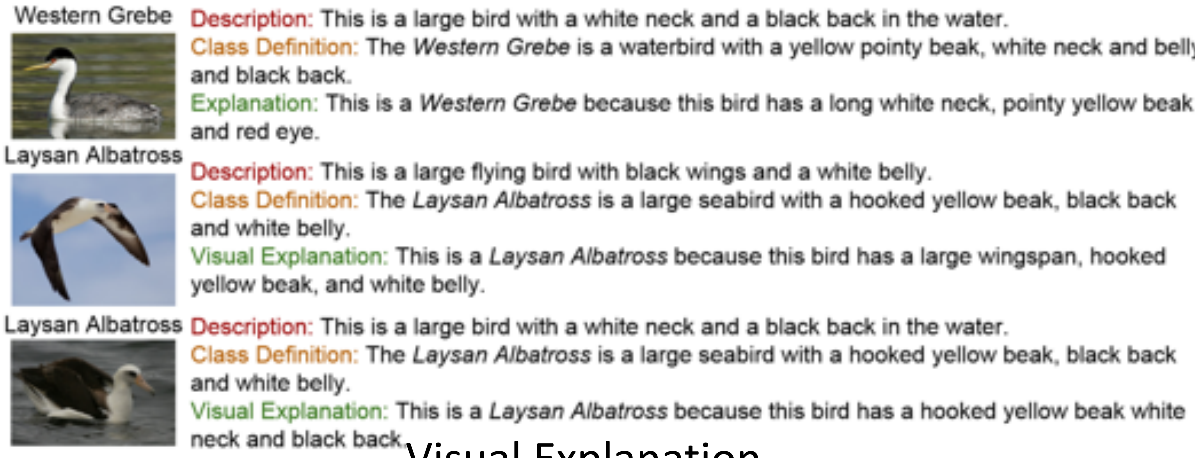
Overview of explanation in different AI fields (2)

- Computer Vision



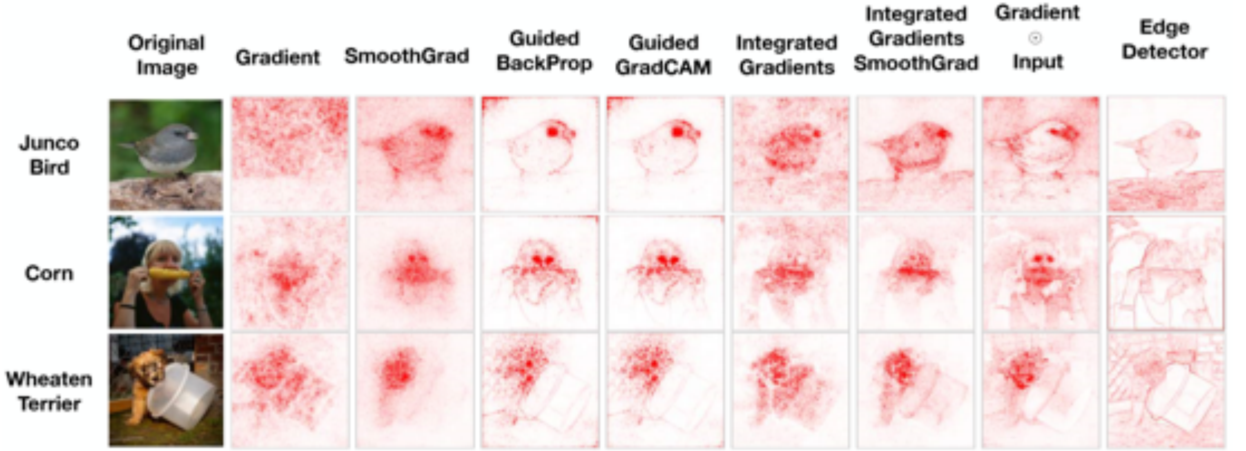
Uncertainty Map

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



Visual Explanation

Lisa Anne Hendricks, Zeynep Akata, Marcus Rohrbach, Jeff Donahue, Bernt Schiele, Trevor Darrell: Generating Visual Explanations. ECCV (4) 2016: 3-19

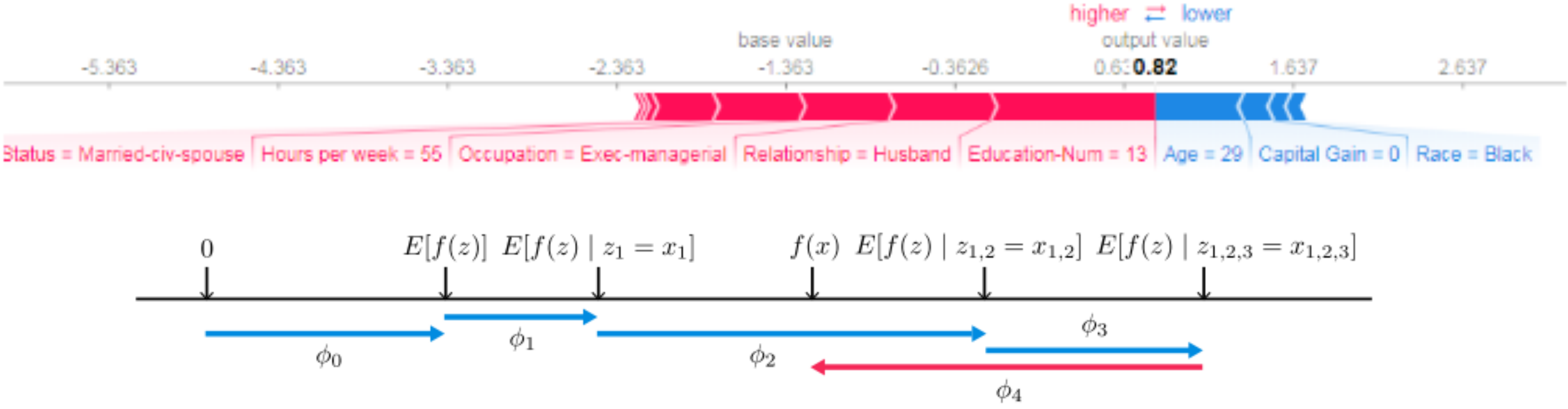


Saliency Map

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

Overview of explanation in different AI fields (3)

- Game Theory

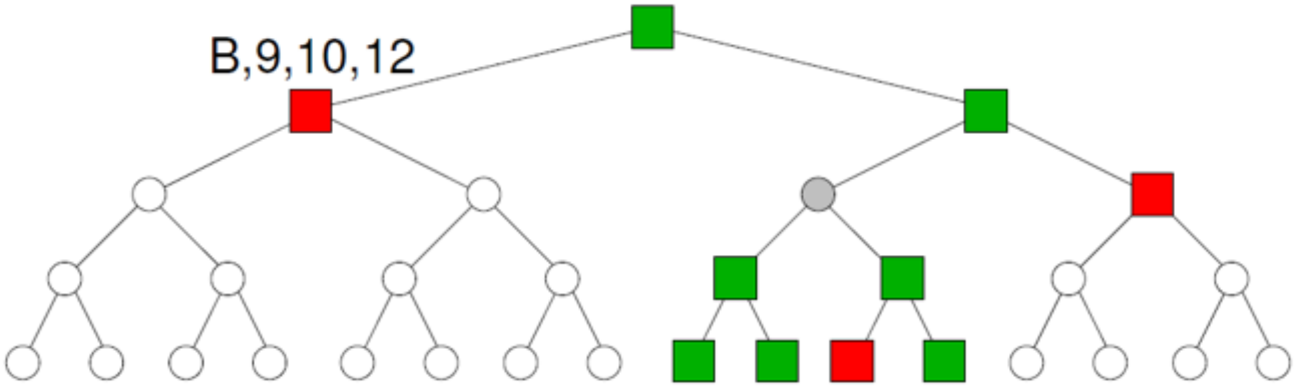


Shapley Additive Explanation

Scott M. Lundberg, Su-In Lee: A Unified Approach to Interpreting Model Predictions. NIPS 2017: 4768-4777

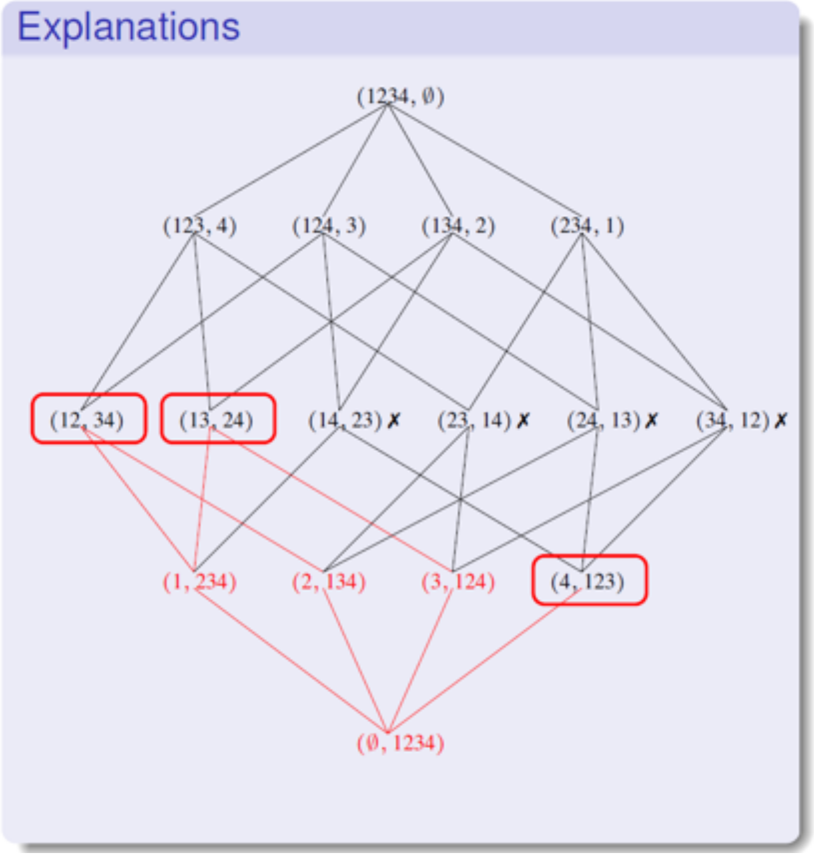
Overview of explanation in different AI fields (4)

- Search and Constraint Satisfaction



Conflicts resolution

Barry O'Sullivan, Alexandre Papadopoulos, Boi Faltings, Pearl Pu: Representative Explanations for Over-Constrained Problems. AAAI 2007: 323-328

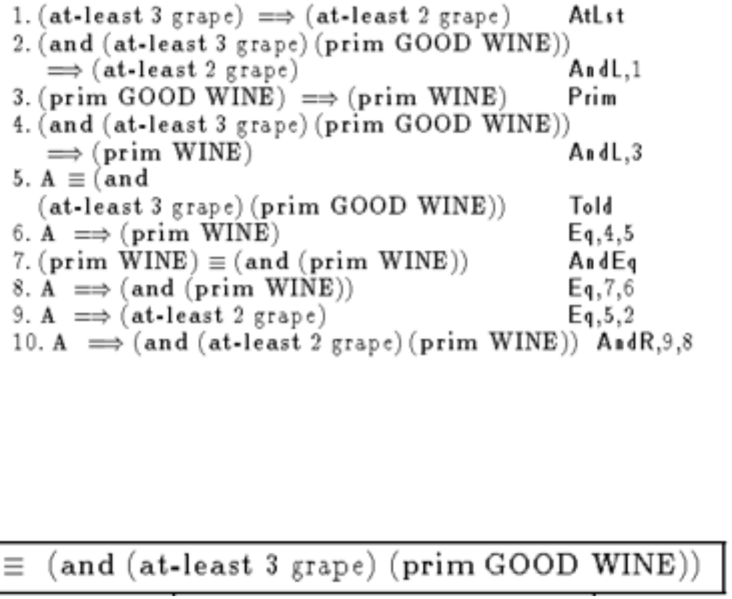


Constraints relaxation

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

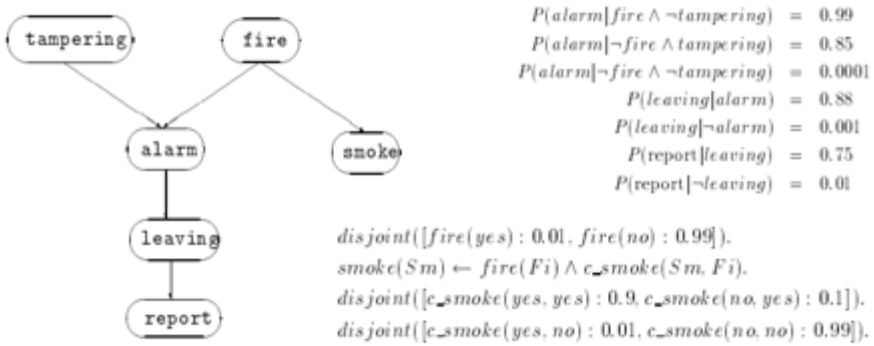
Overview of explanation in different AI fields (5)

- Knowledge Representation and Reasoning



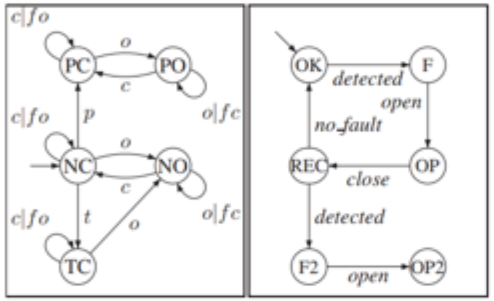
Explaining Reasoning (through Justification) e.g., Subsumption

Deborah L. McGuinness, Alexander Borgida: Explaining Subsumption in Description Logics. IJCAI (1) 1995: 816-821



Abduction Reasoning (in Bayesian Network)

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

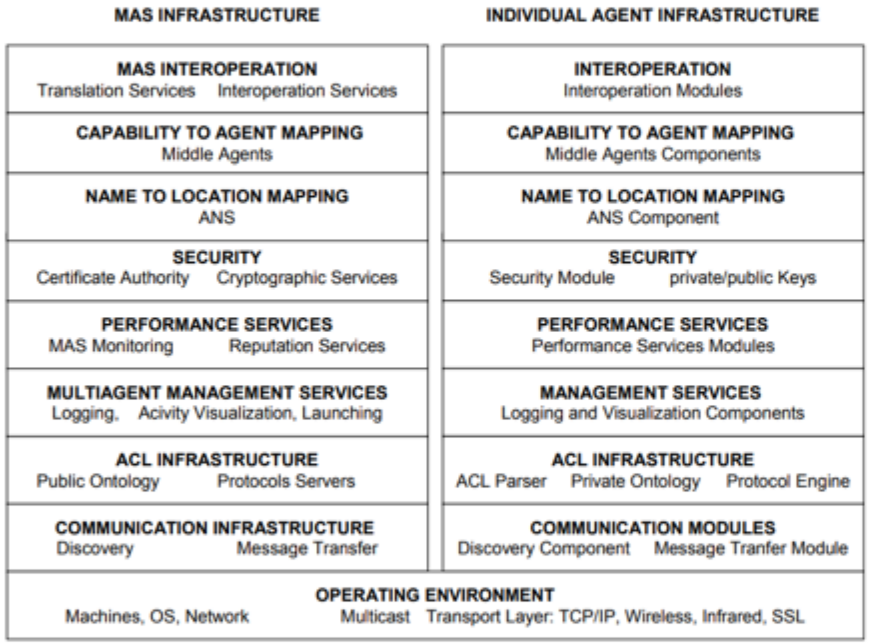


Diagnosis Inference

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

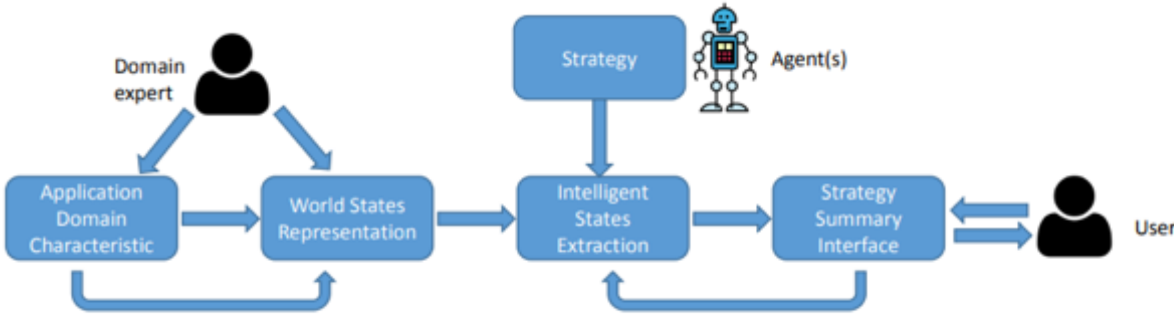
Overview of explanation in different AI fields (6)

- Multi-agent Systems



Explanation of Agent Conflicts and Harmful Interactions

Katia P. Sycara, Massimo Paolucci, Martin Van Velsen, Joseph A. Giampapa: The RETSINA MAS Infrastructure. Autonomous Agents and Multi-Agent Systems 7(1-2): 29-48 (2003)



Agent Strategy Summarization

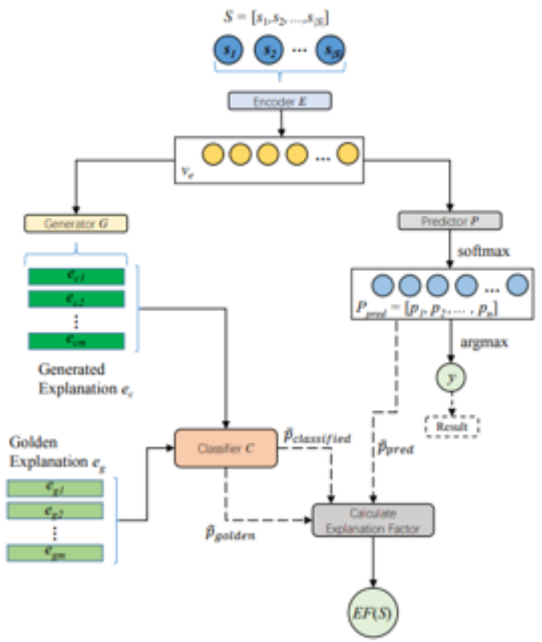
Ofra Amir, Finale Doshi-Velez, David Sarne: Agent Strategy Summarization. AAMAS 2018: 1203-1207

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

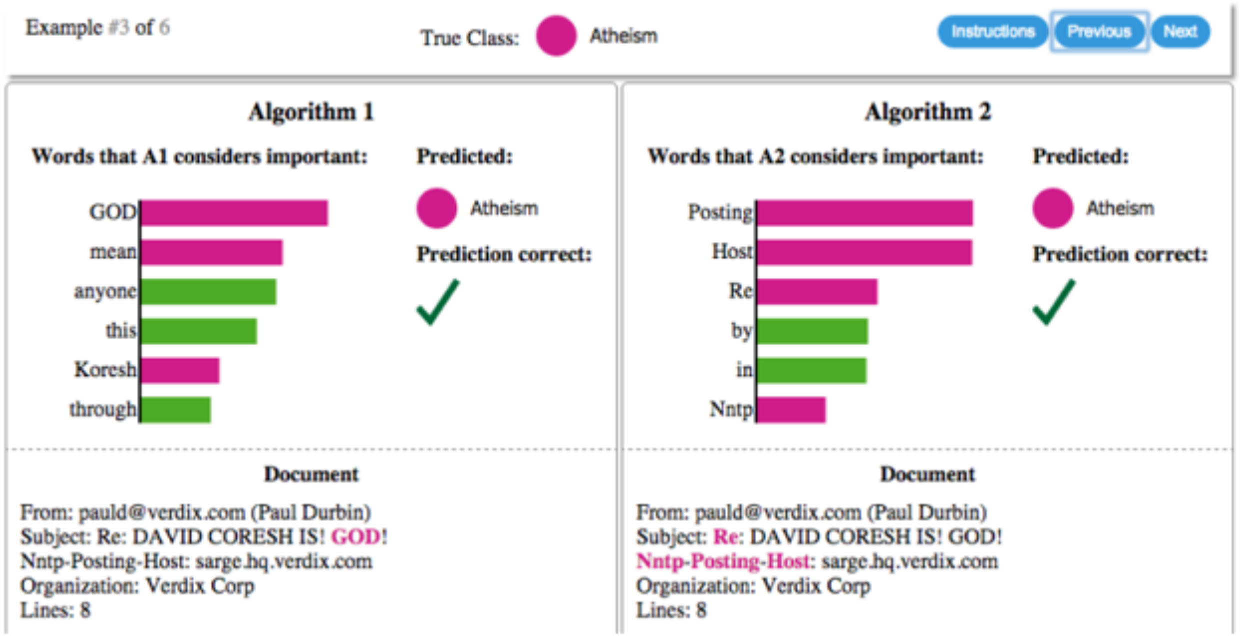
Overview of explanation in different AI fields (7)

- NLP



Fine-grained explanations are in the form of:

- texts in a real-world dataset;
- Numerical scores



Explainable NLP

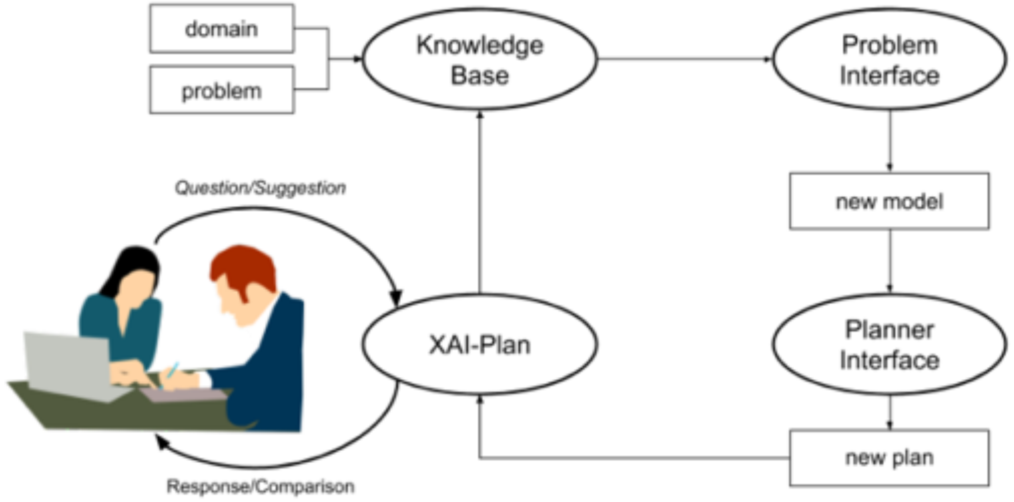
LIME for NLP

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

Marco Túlio Ribeiro, Sameer Singh, Carlos Guestrin: "Why Should I Trust You?": Explaining the Predictions of Any Classifier. KDD 2016: 1135-1144

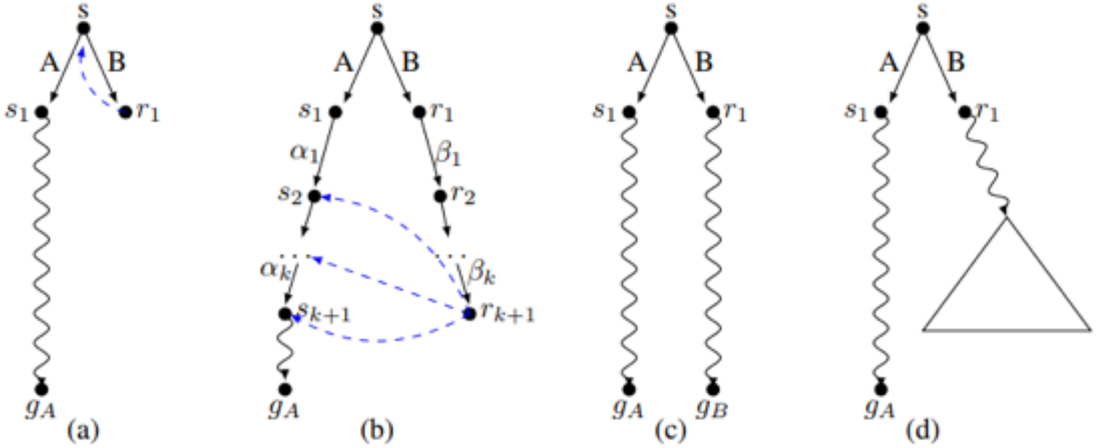
Overview of explanation in different AI fields (8)

- Planning and Scheduling



XAI Plan

Rita Borgo, Michael Cashmore, Daniele Magazzeni: Towards Providing Explanations for AI Planner Decisions. CoRR abs/1810.06338 (2018)



Human-in-the-loop Planning

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

Overview of explanation in different AI fields (9)

- Robotics

Specificity, S		Abstraction, A			
		Level 1	Level 2	Level 3	Level 4
	General Picture	Start and finish point of the complete route	Total distance and time taken for the complete route	Total distance and time taken for the complete route	Starting and ending landmark of complete route
	Summary	Start and finish point for subroute on each floor of each building	Total distance and time taken for subroute on each floor of each building	Total distance and angles for subroute on each floor of each building	Starting and ending landmark for subroute on each floor of each building
	Detailed Narrative	Start and finish points of complete route plus time taken for each edge of route	Angle turned at each point plus the total distance and time taken for each edge of route	Turn direction at each point plus total distance for each edge of route	All landmarks encountered on the route

Narration of Autonomous Robot Experience

Stephanie Rosenthal, Sai P Selvaraj, and Manuela Veloso. Verbalization: Narration of autonomous robot experience. In IJCAI, pages 862–868. AAAI Press, 2016.

Daniel J Brooks et al. 2010. Towards State Summarization for Autonomous Robots.. In AAAI Fall Symposium: Dialog with Robots, Vol. 61. 62.

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

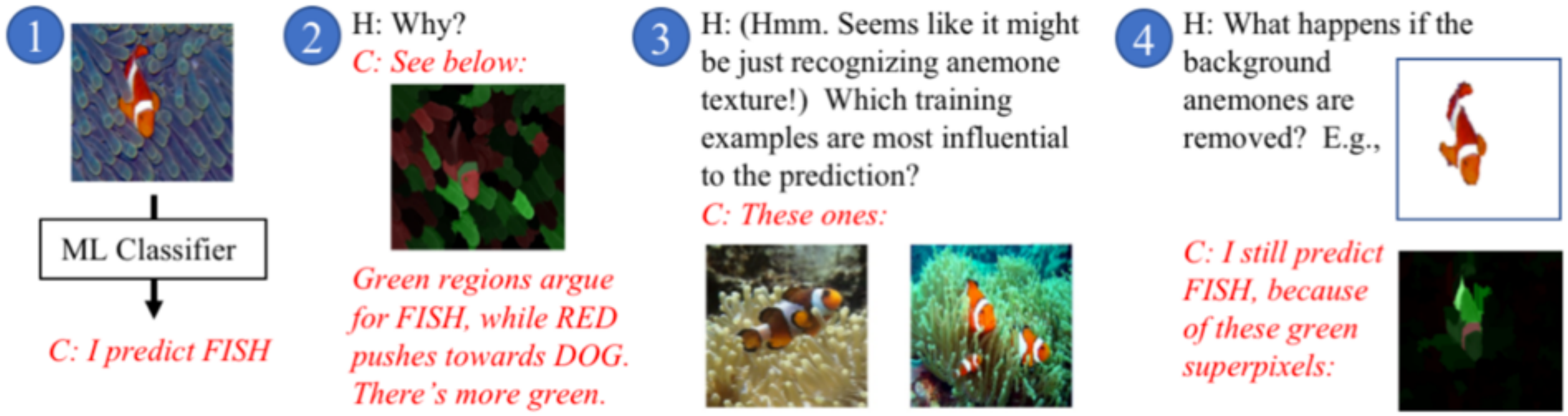
Raymond Ka-Man Sheh: "Why Did You Do That?" Explainable Intelligent Robots. AAAI Workshops 2017

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


Vous préférez
un conseiller
qui répond
humainement

ou une machine
qui répond
machinalement ?

Oxford Dictionary of English

explanation | ɛkspləˈneɪʃ(ə)n |

noun

a statement or account that makes something clear: *the birth rate is central to any explanation of population trends*.

interpret | ɪnˈtəːprɪt |

verb **(interprets, interpreting, interpreted)** *[with object]*

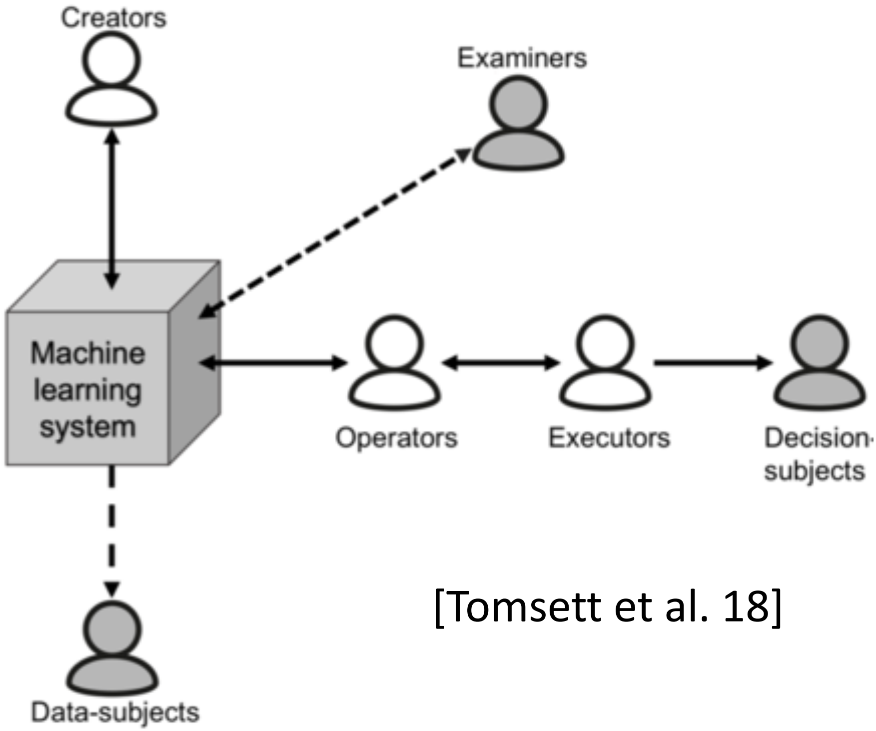
1 explain the meaning of (information or actions): *the evidence is difficult to interpret*.

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?"



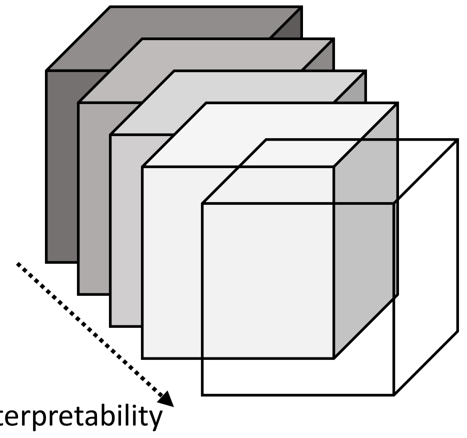
[Tomsett et al. 18]

An ideal explainer should model the *user background.*

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

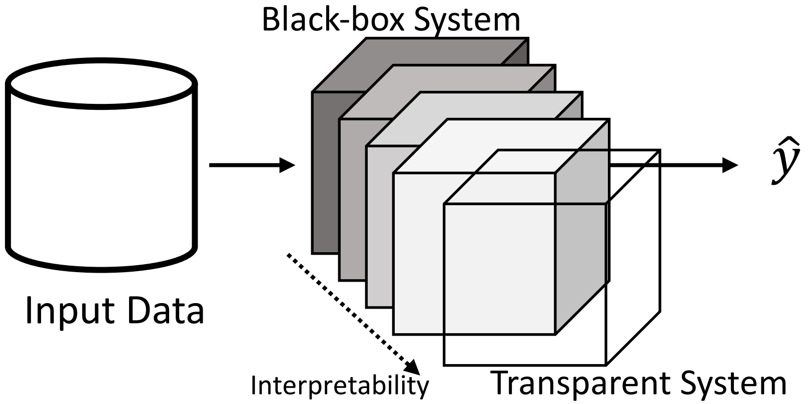
Evaluation: Interpretability as Latent Property

- 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?

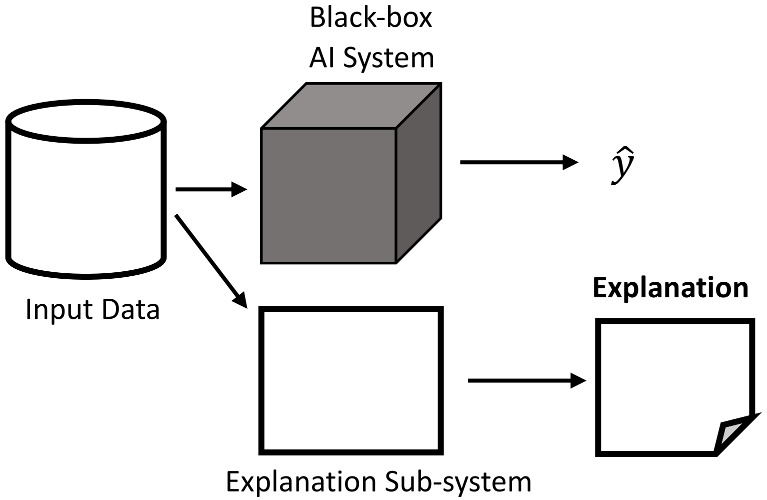


Explainable AI Systems

Transparent-by-design systems



Post-hoc Explanation (black-box explanation) systems



[Mittelstadt et al. 2018]

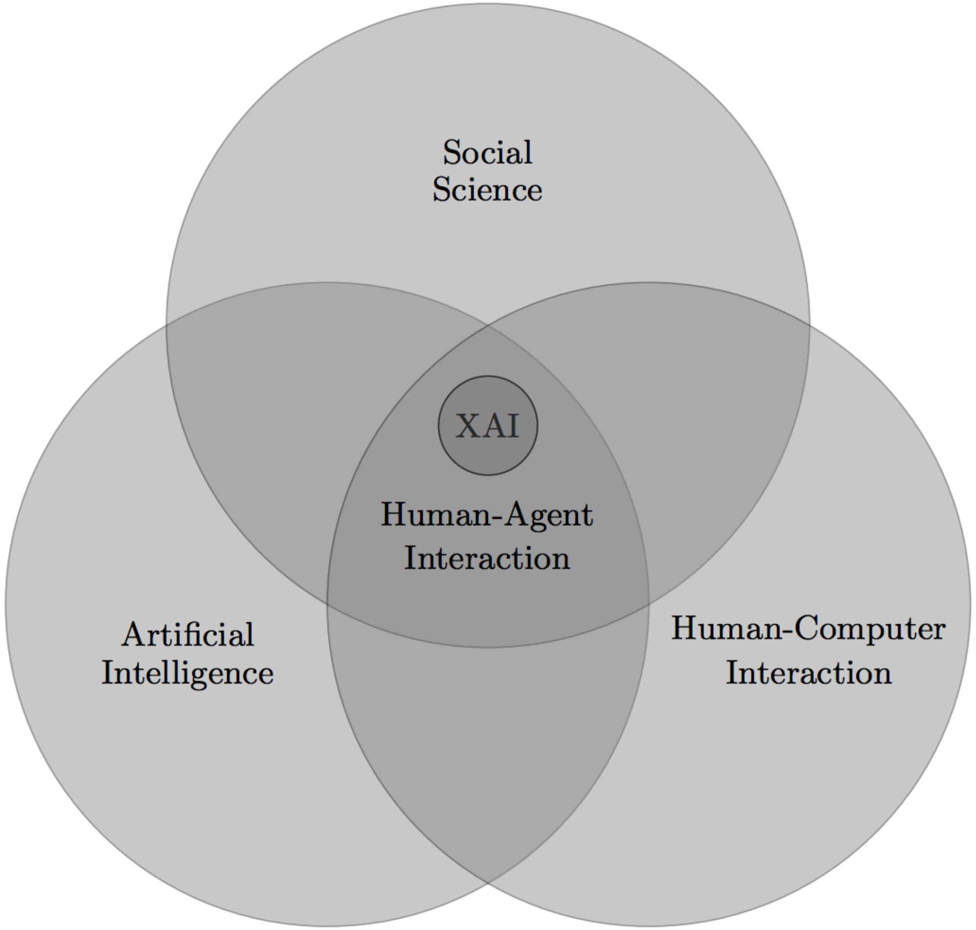
(Some) Desired Properties of Explainable AI Systems

- 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.

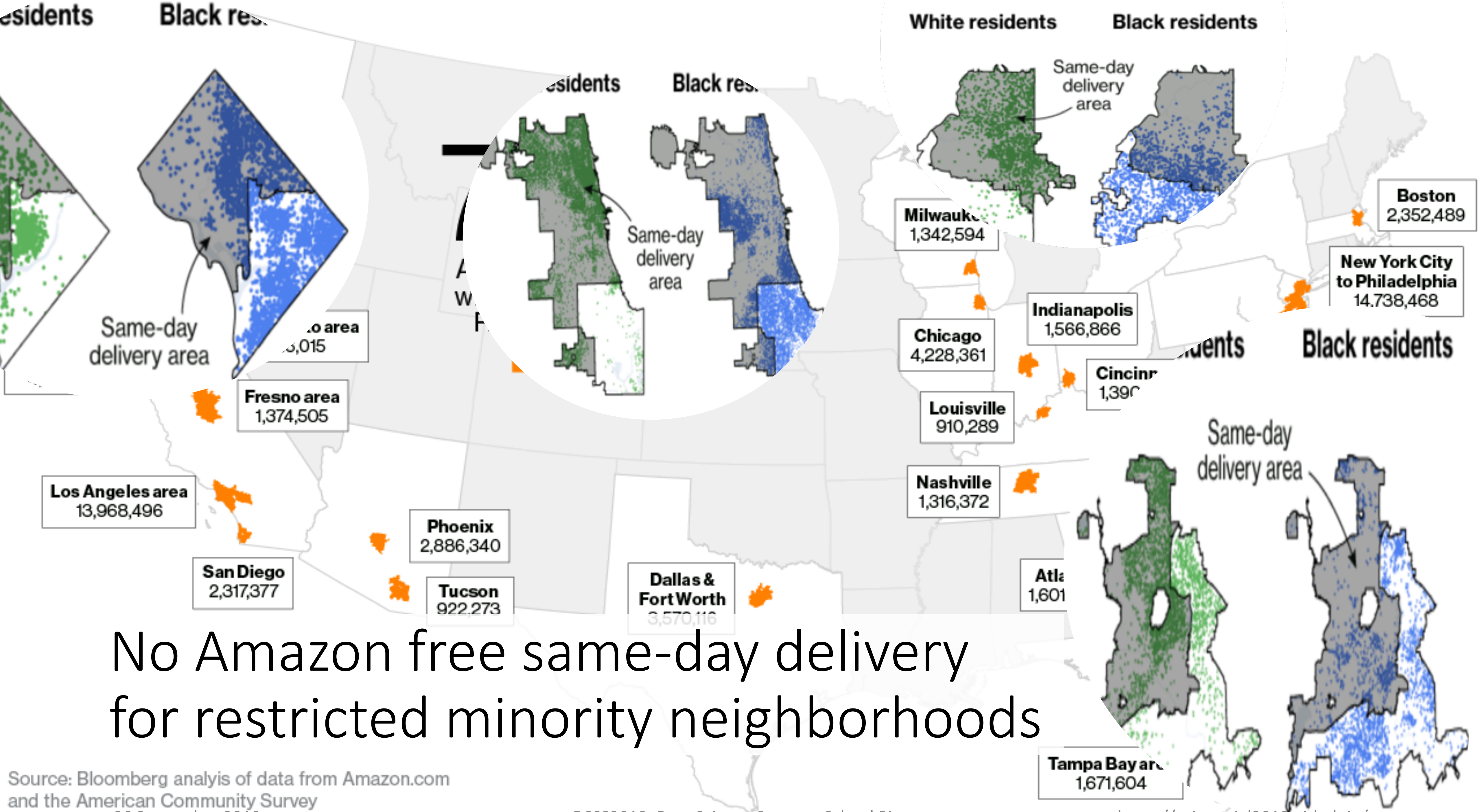
[Weld and Bansal 2018] Weld, D., and Gagan Bansal. "The challenge of crafting intelligible intelligence." Communications of ACM (2018).

[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

Bias in Machine Learning

COMPAS recidivism black bias



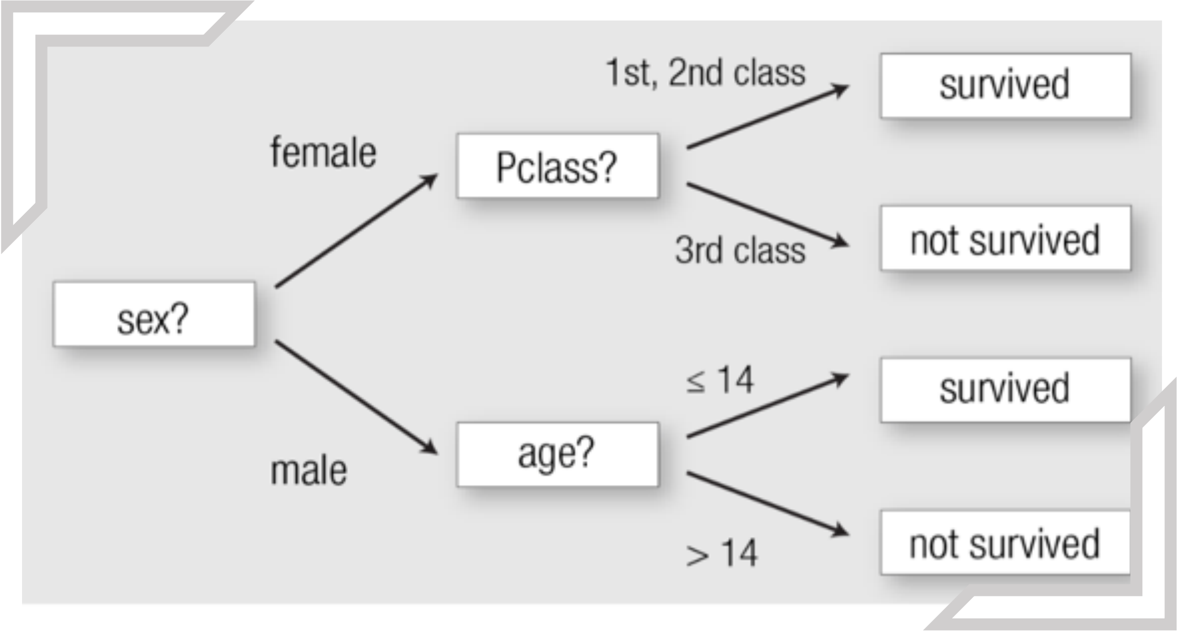
No Amazon free same-day delivery for restricted minority neighborhoods

The background bias

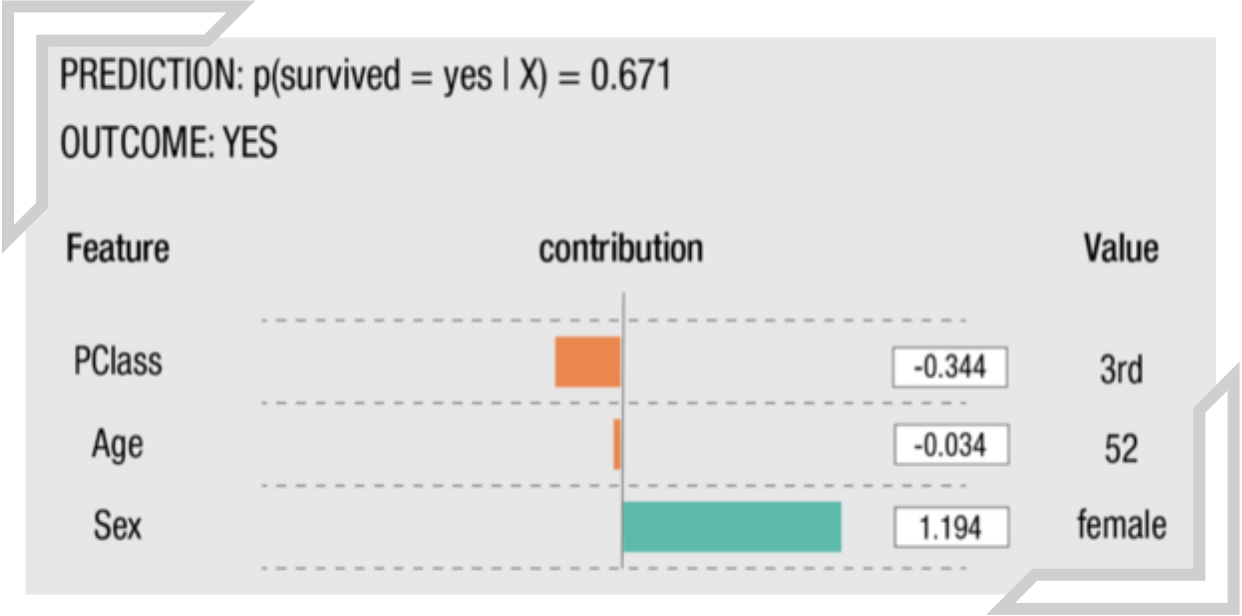
(a) Husky classified as wolf (b) Explanation

Interpretable ML Models

Recognized Interpretable Models



Decision Tree

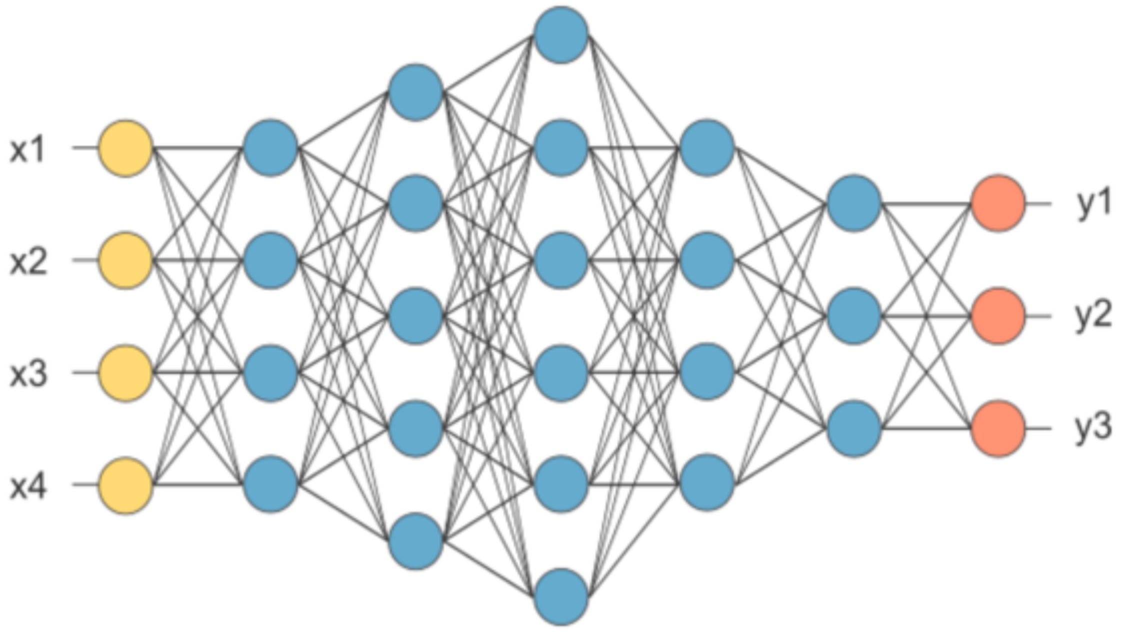


Linear Model

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

Rules

Black Box Model



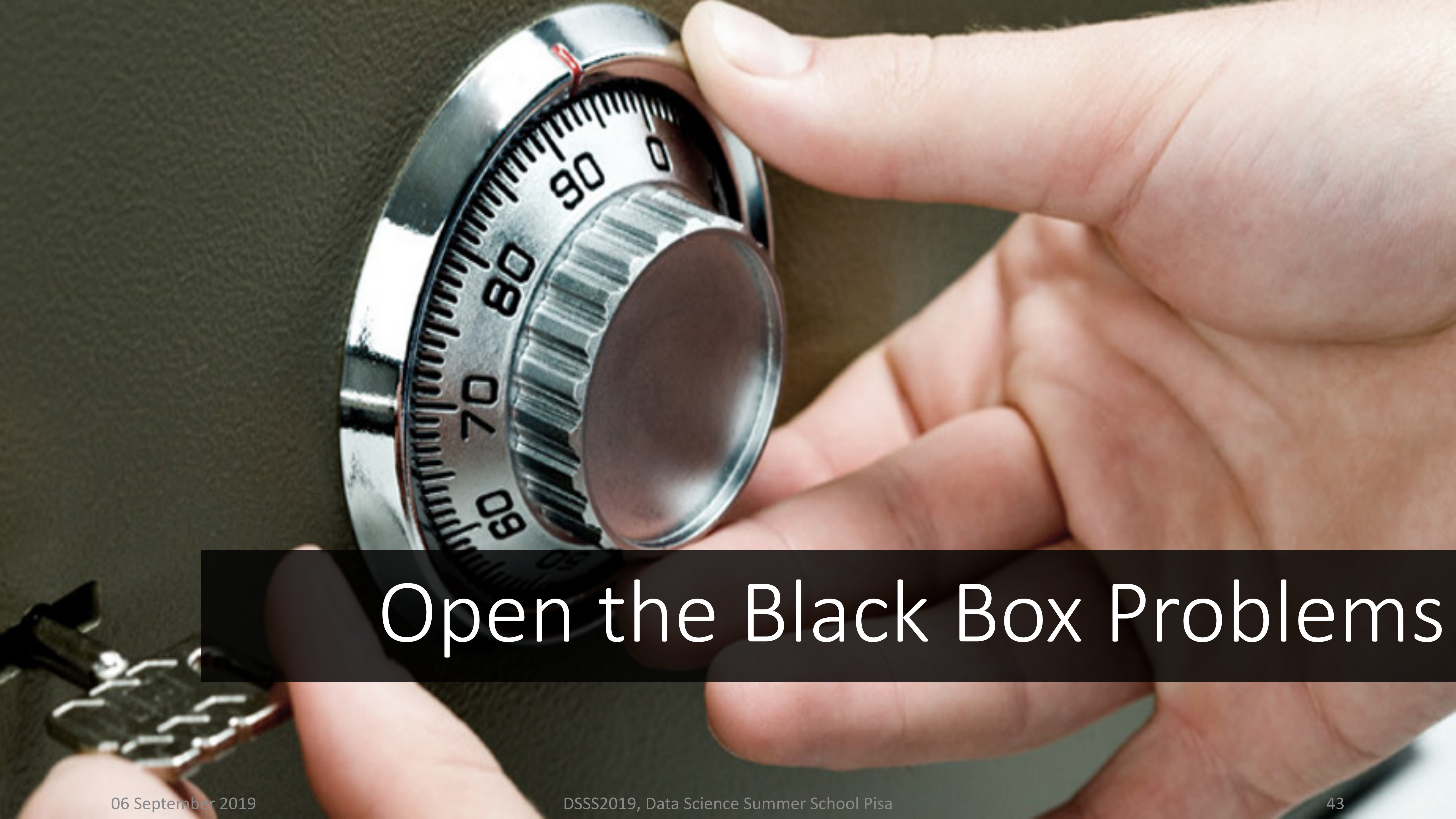
A ***black box*** is a DMML model, whose internals are either unknown to the observer or they are known but uninterpretable by humans.

- 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.

Complexity

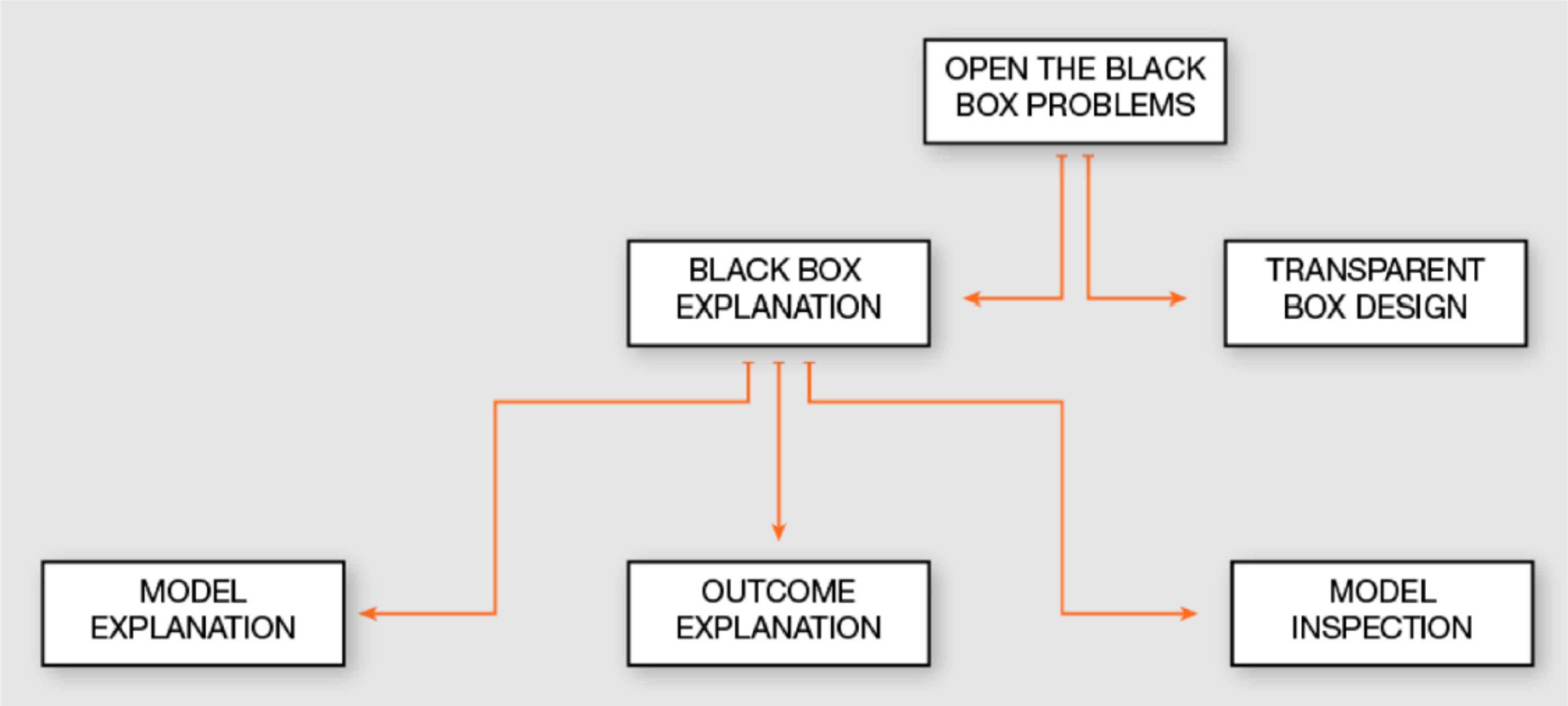
- Opposed to *interpretability*.
- Is only related to the model and not to the training data that is unknown.
- Generally estimated with a rough approximation related to the ***size*** of the interpretable model.
- Linear Model: number of non zero weights in the model.
- Rule: number of attribute-value 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.

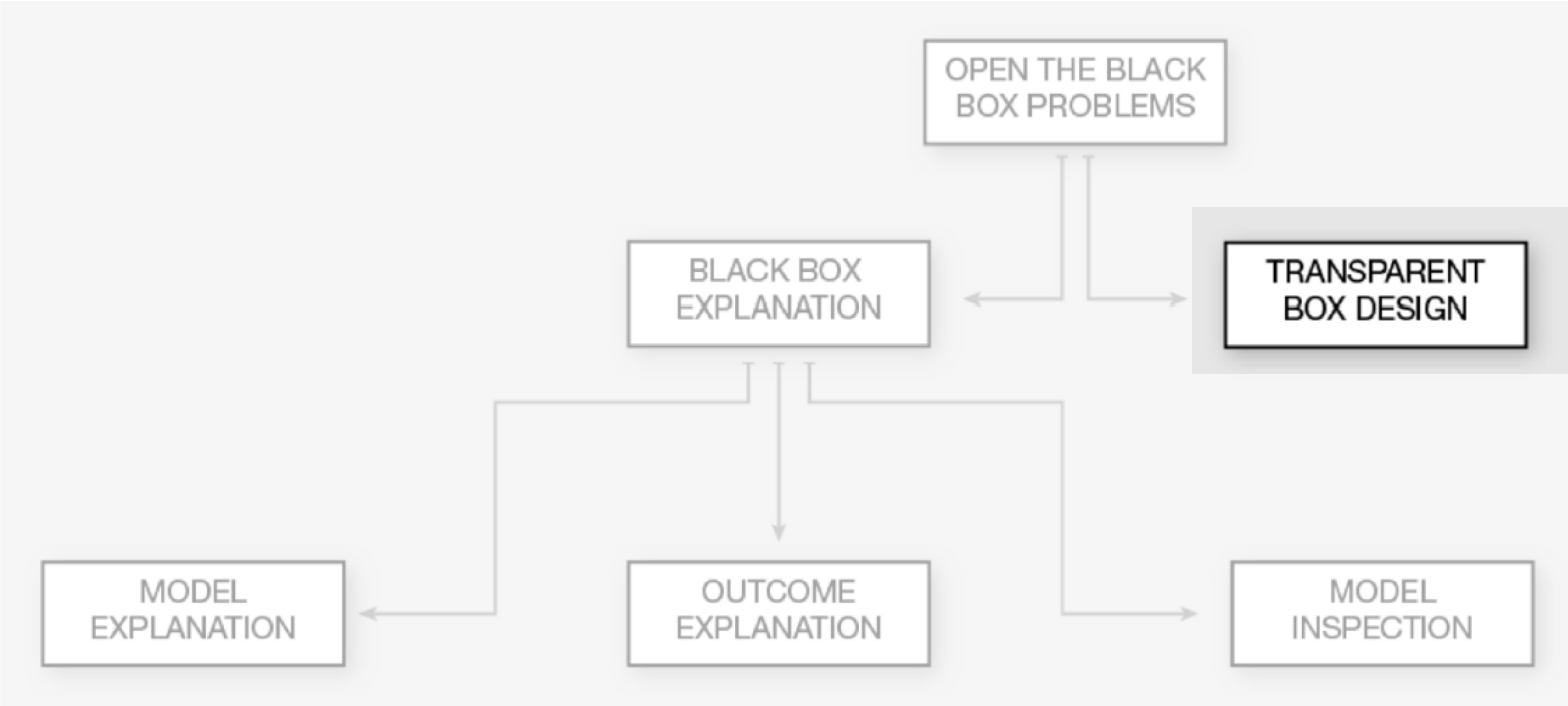


Open the Black Box Problems

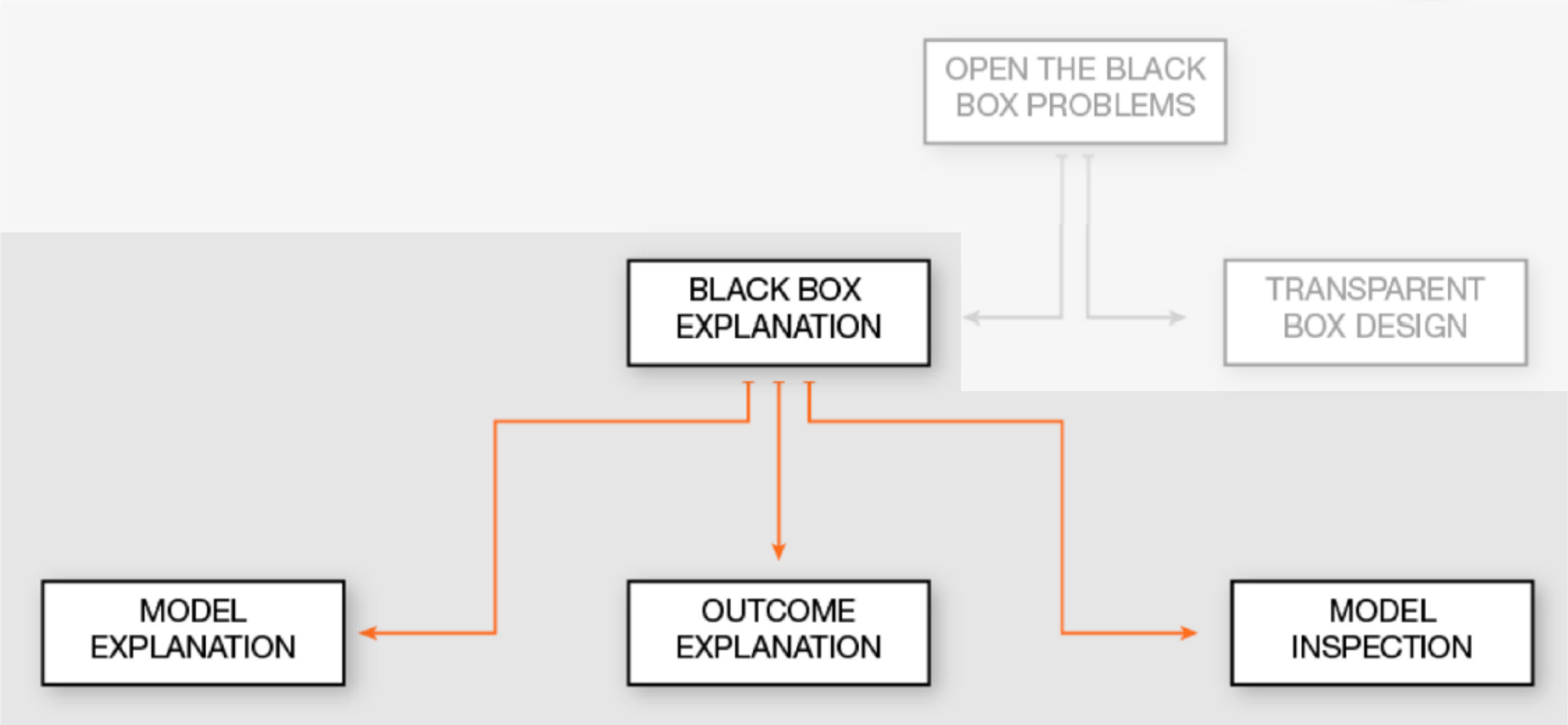
Problems Taxonomy



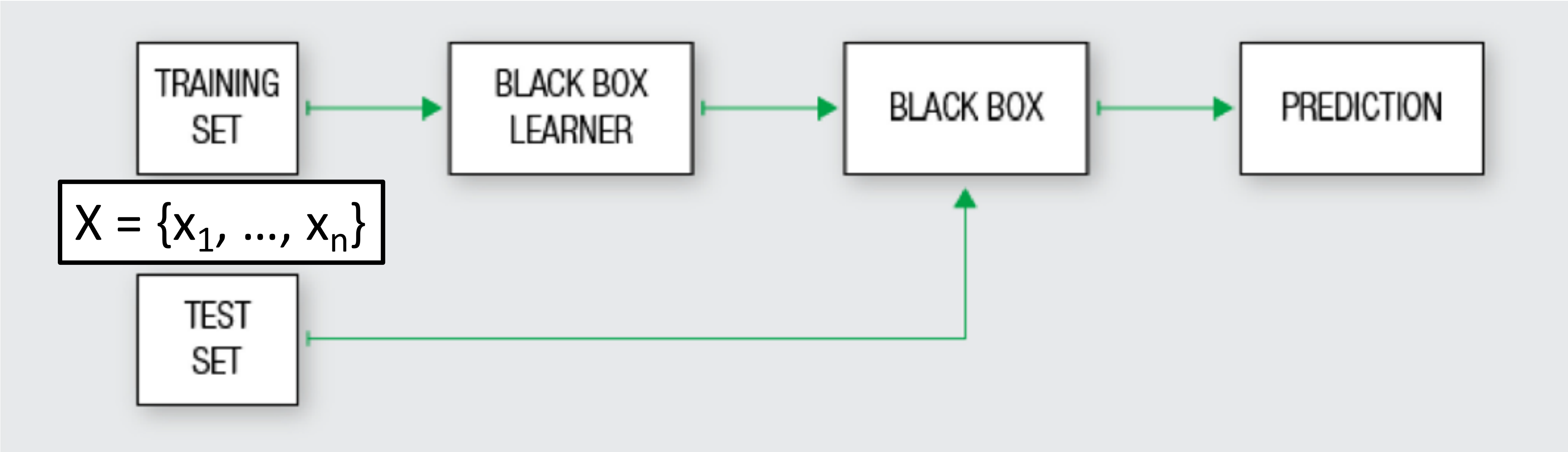
XbD – eXplanation by Design



BBX - Black Box eXplanation

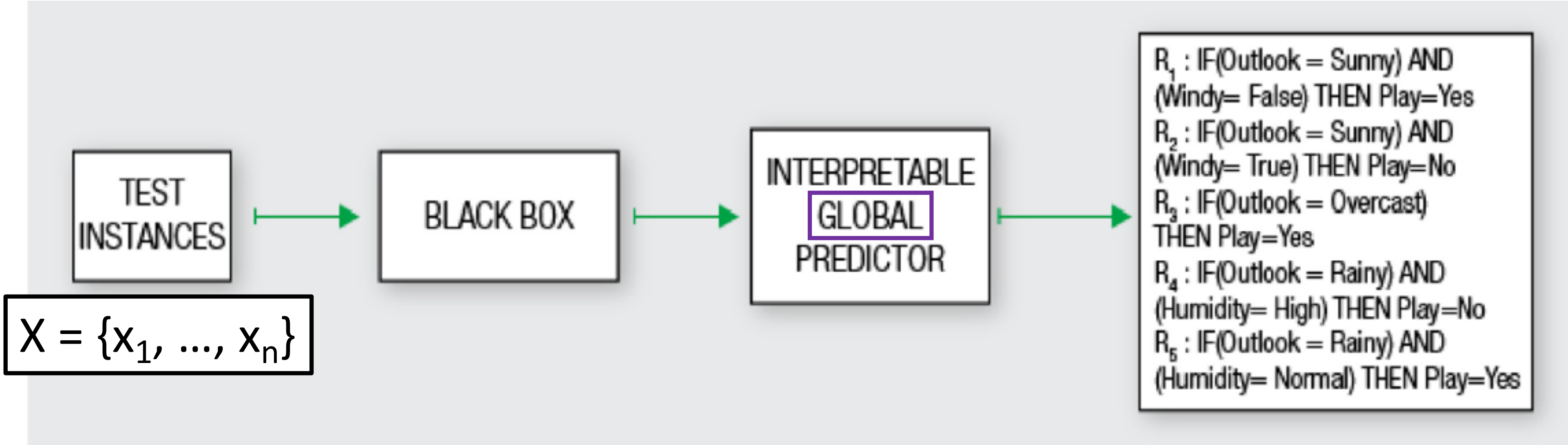


Classification Problem



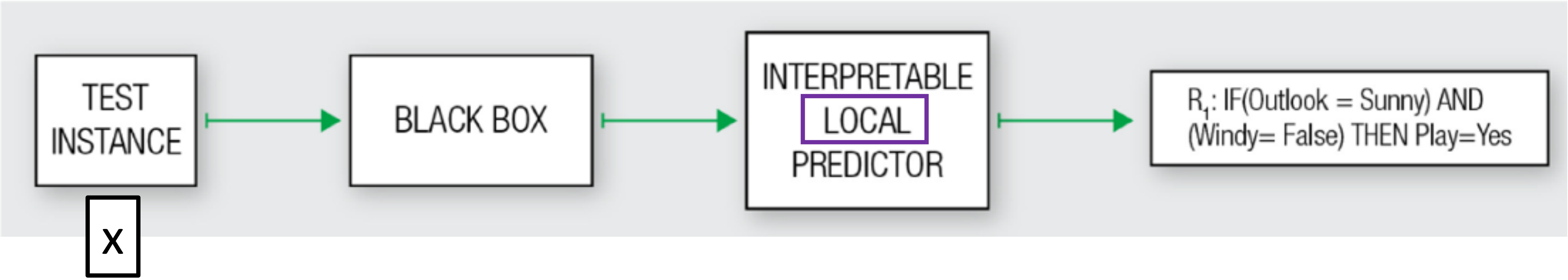
Model Explanation Problem

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



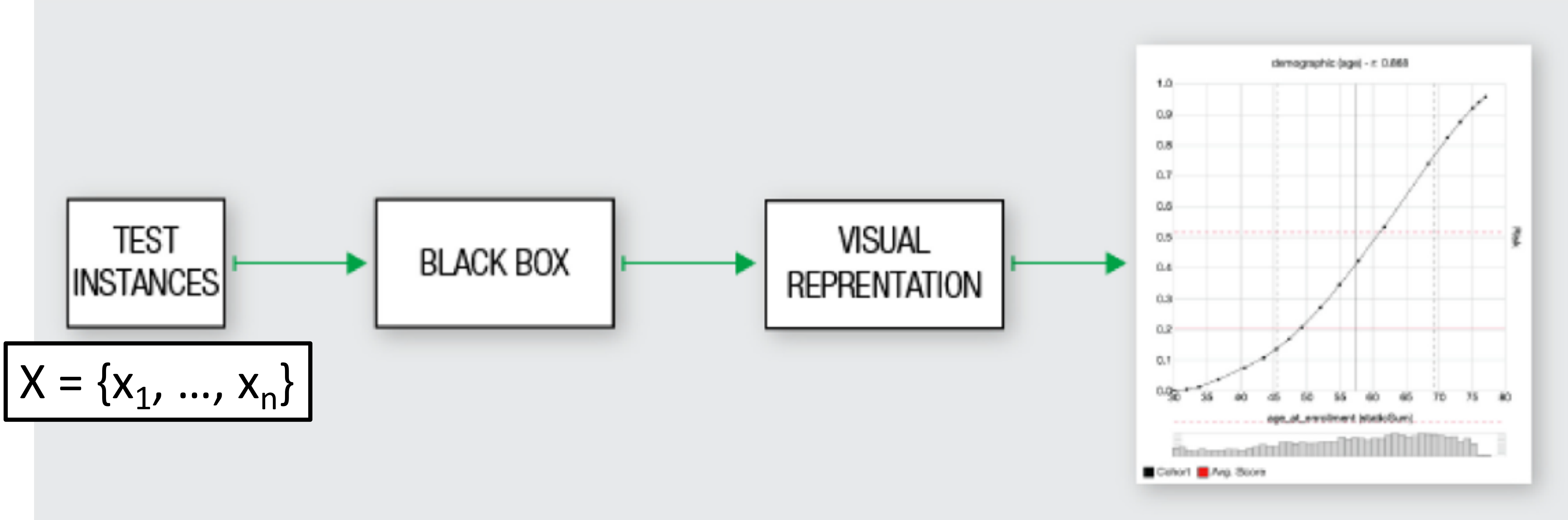
Outcome Explanation Problem

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



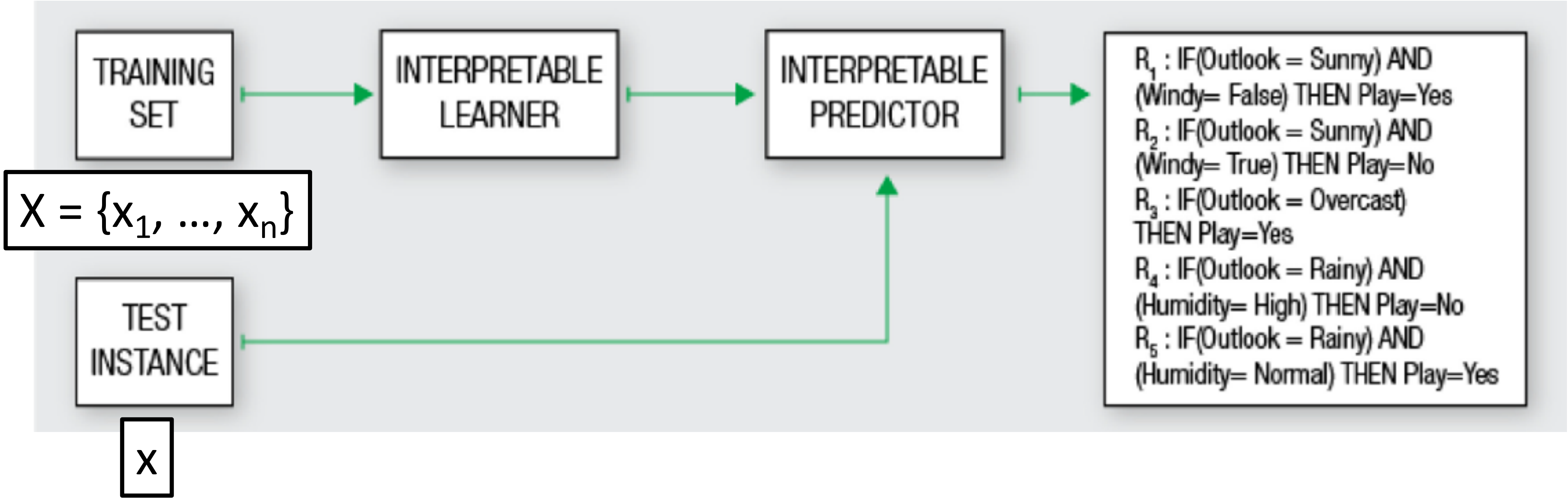
Model Inspection Problem

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.

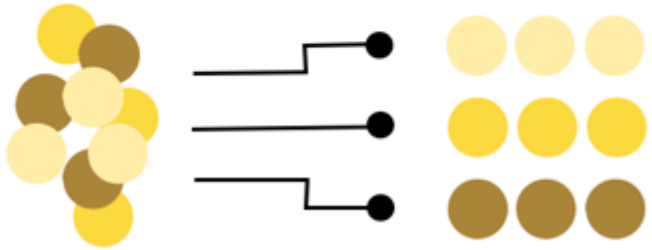


Transparent Box Design Problem

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

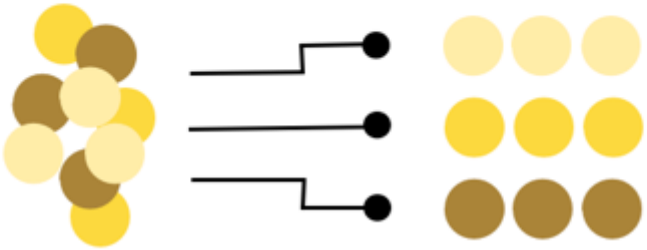


Categorization



- 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

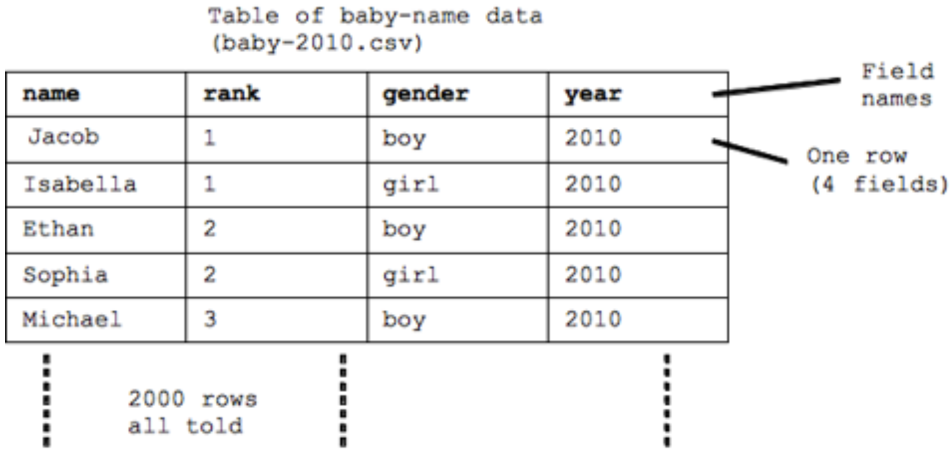


Table of baby-name data
(baby-2010.csv)

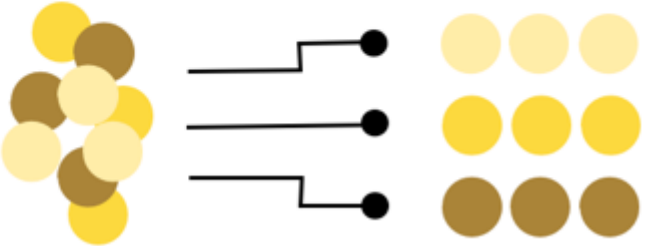
name	rank	gender	year
Jacob	1	boy	2010
Isabella	1	girl	2010
Ethan	2	boy	2010
Sophia	2	girl	2010
Michael	3	boy	2010

Tabular

(TAB)

Images

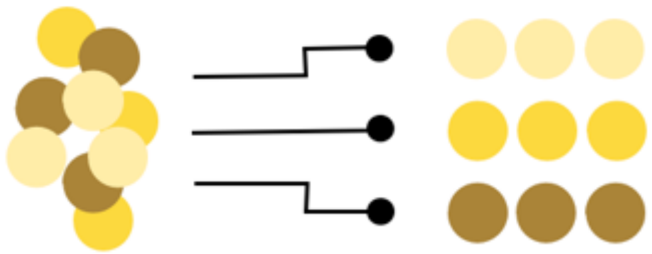
(IMG)



Text

(TXT)

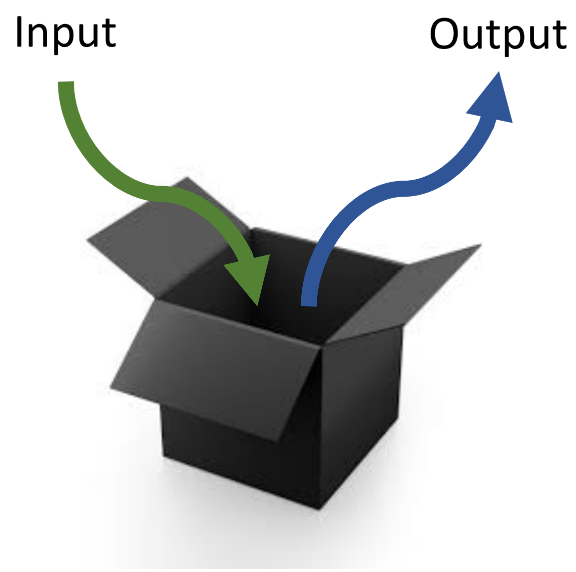
Explanators



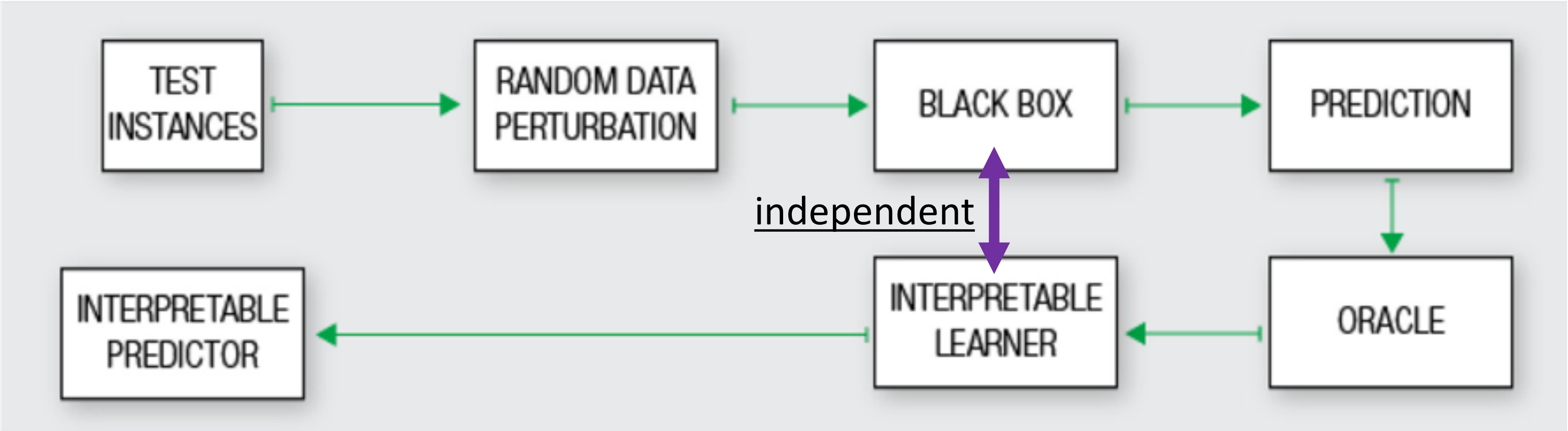
- Decision Tree (***DT***)
- Decision Rules (***DR***)
- Features Importance (***FI***)
- Saliency Mask (***SM***)
- Sensitivity Analysis (***SA***)
- Partial Dependence Plot (***PDP***)
- Prototype Selection (***PS***)
- Activation Maximization (***AM***)

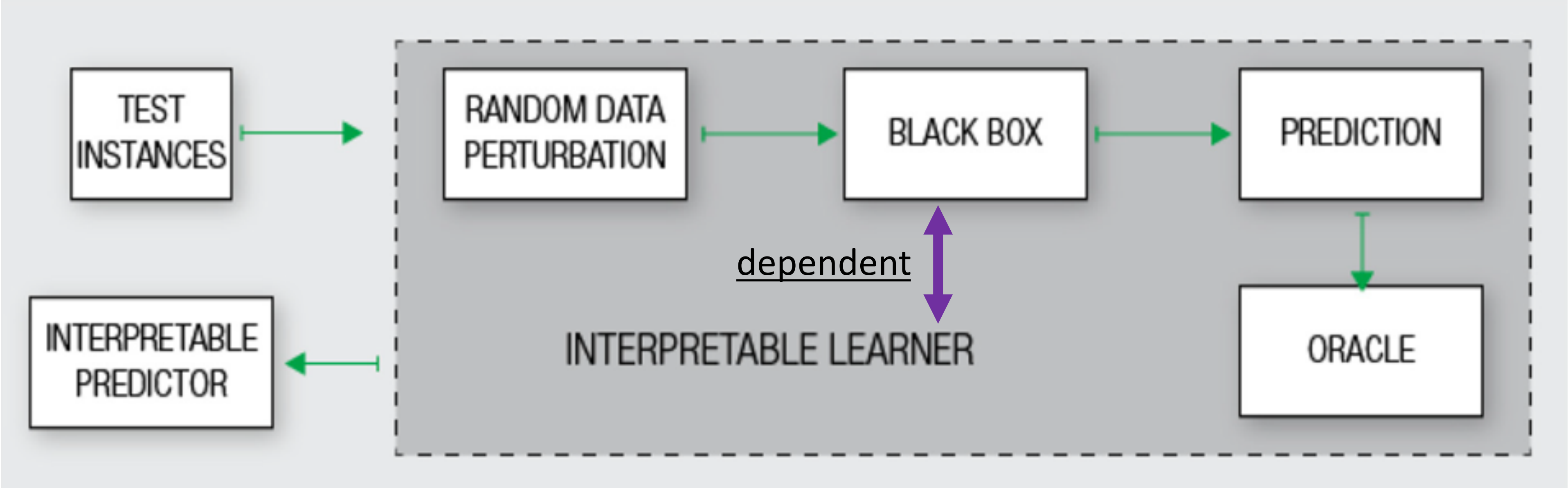
Reverse Engineering

- 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





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

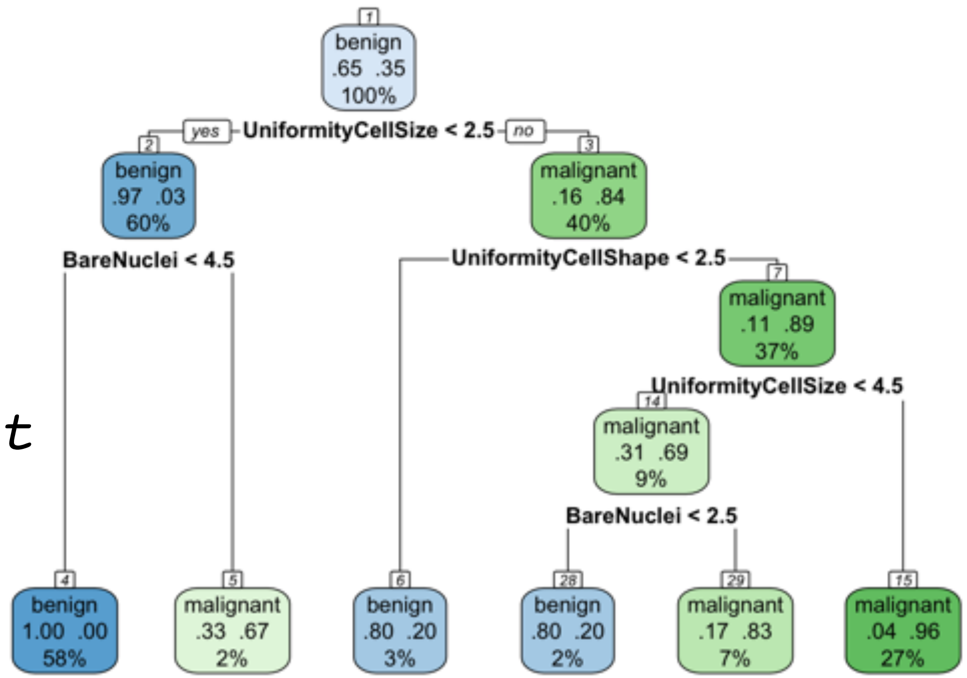
- Explanator: DT
  - Black Box: NN, TE
  - Data Type: TAB
- Explanator: DR
  - Black Box: NN, SVM, TE
  - Data Type: TAB
- Explanator: FI
  - Black Box: AGN
  - Data Type: TAB

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

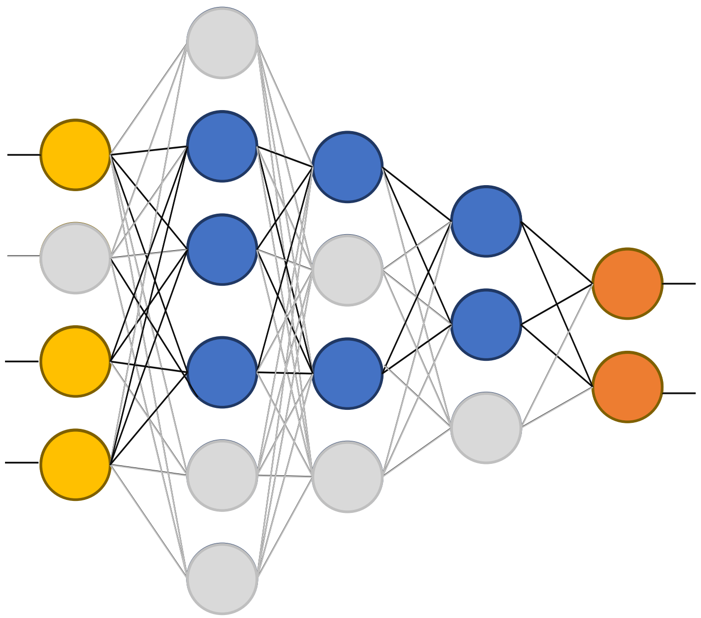
```
T = root_of_the_tree()
Q = <T, X, {}>
while Q not empty & size(T) < limit
      N, X_N, C_N  = pop(Q)
      Z_N = random(X_N, C_N)
      y_Z = b(Z), y = b(X_N)
      if same_class(y U y_Z)
            continue
      S = best_split(X_N U Z_N, y U y_Z)
      S'= best_m-of-n_split(S)
      N = update_with_split(N,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

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

else

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

else

$$\text{class} = C_2$$

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

Solving The Outcome Explanation Problem

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	✓	✓	✓		✓

Local Model Explainers

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

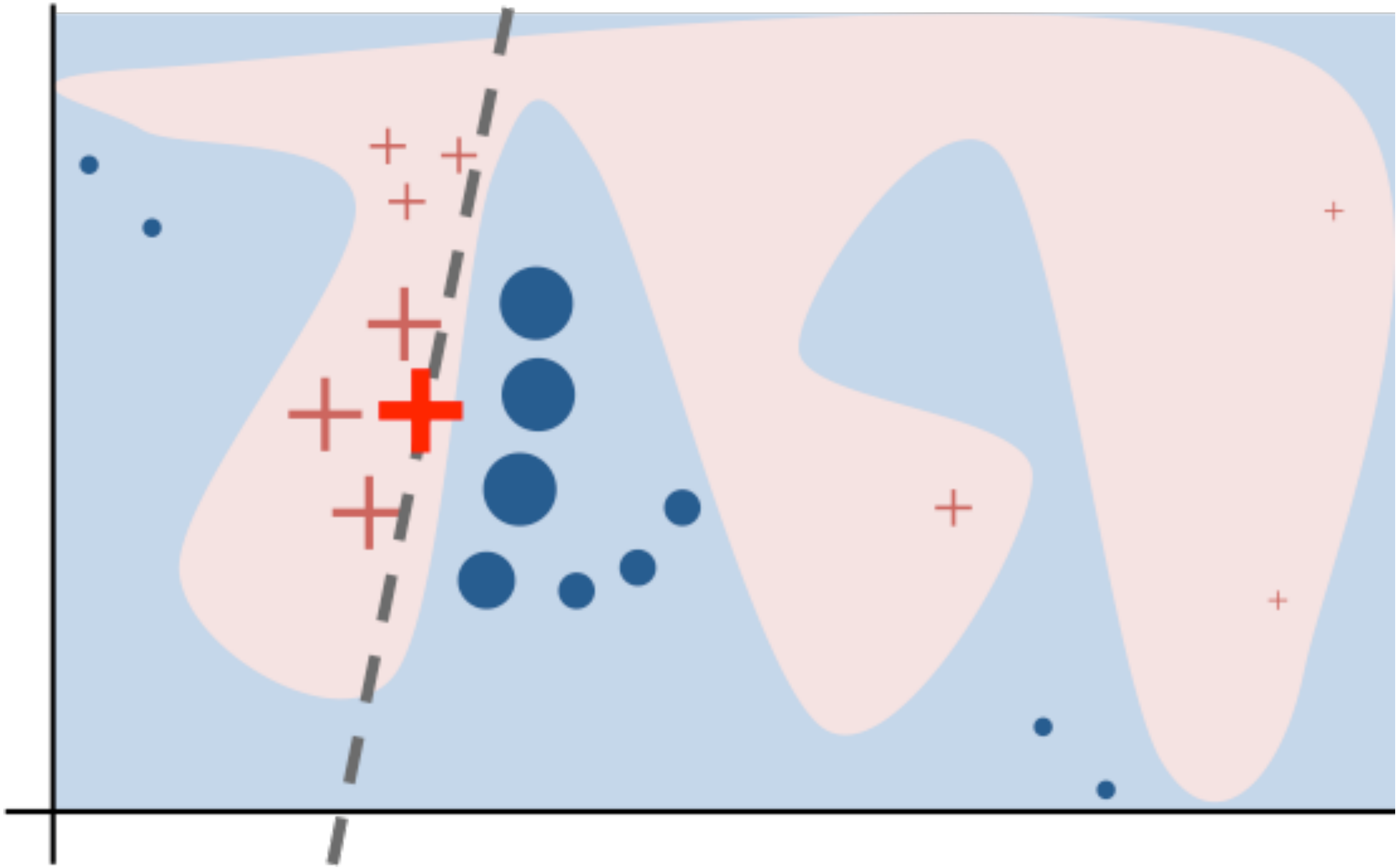
- Explanator: FI
  - Black Box: DNN, SVM
  - Data Type: ANY

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

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

Local Explanation

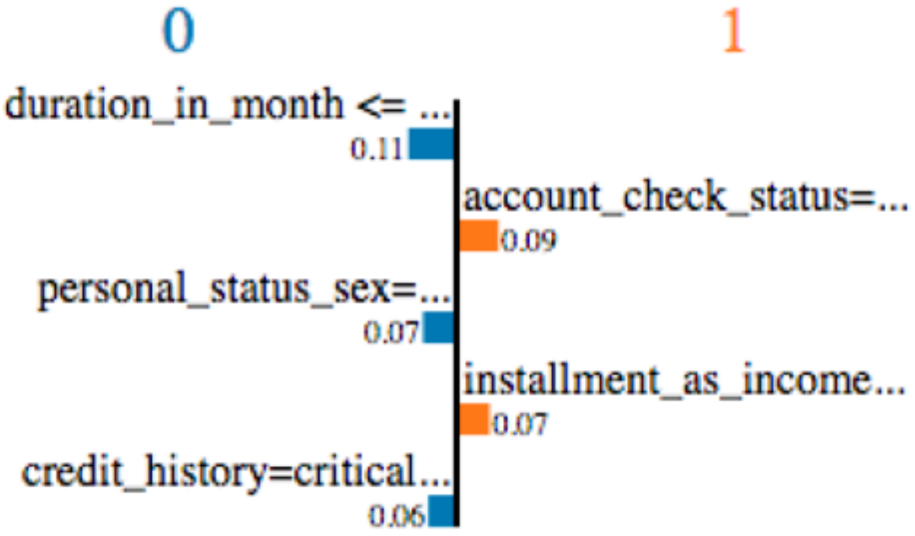
- The overall decision boundary is complex
- In the neighborhood of a single decision, the boundary is simple
- A single decision can be explained by auditing the black box around the given instance and learning a ***local*** decision.



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 U {<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.

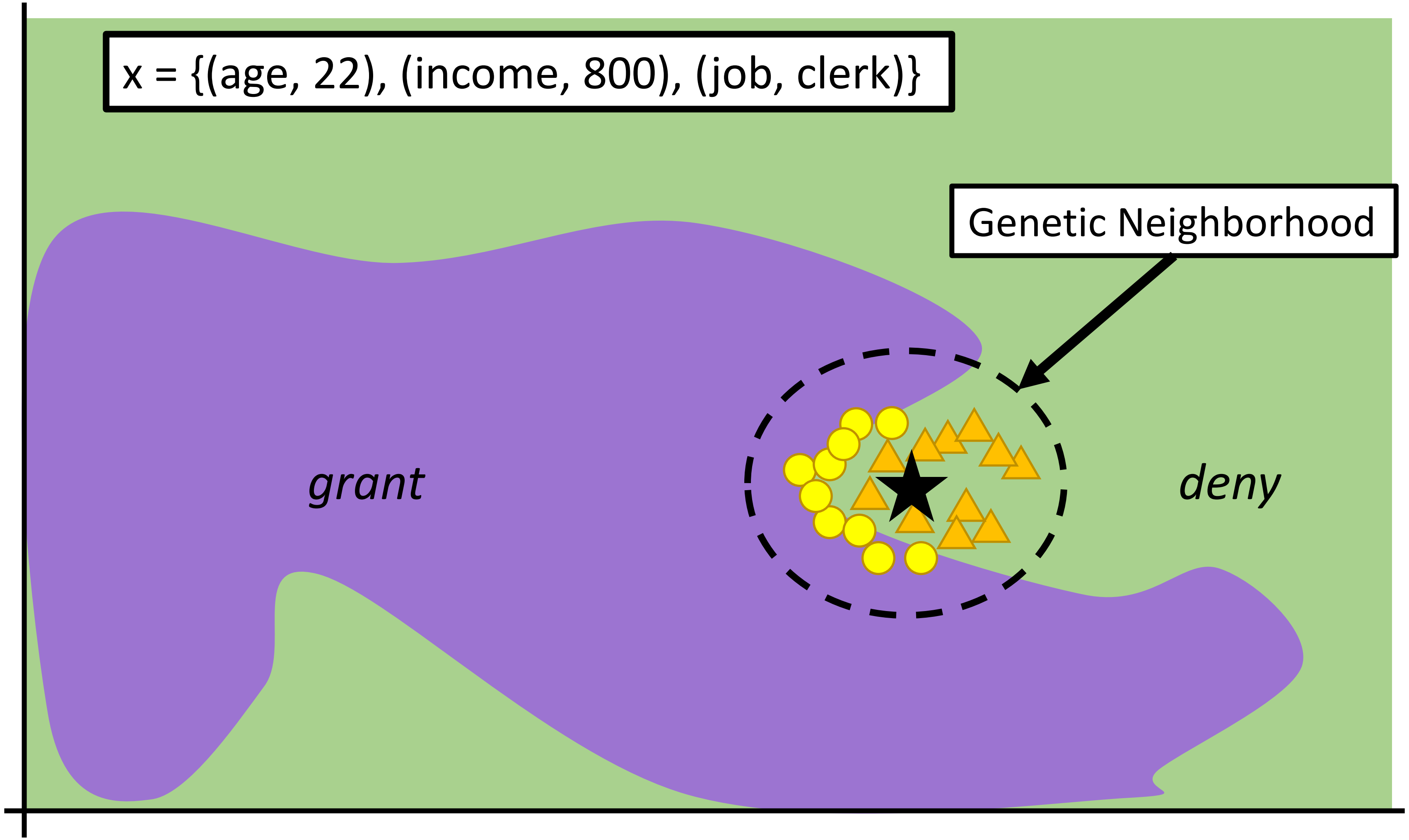
LORE – DR, AGN, TAB

```
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.

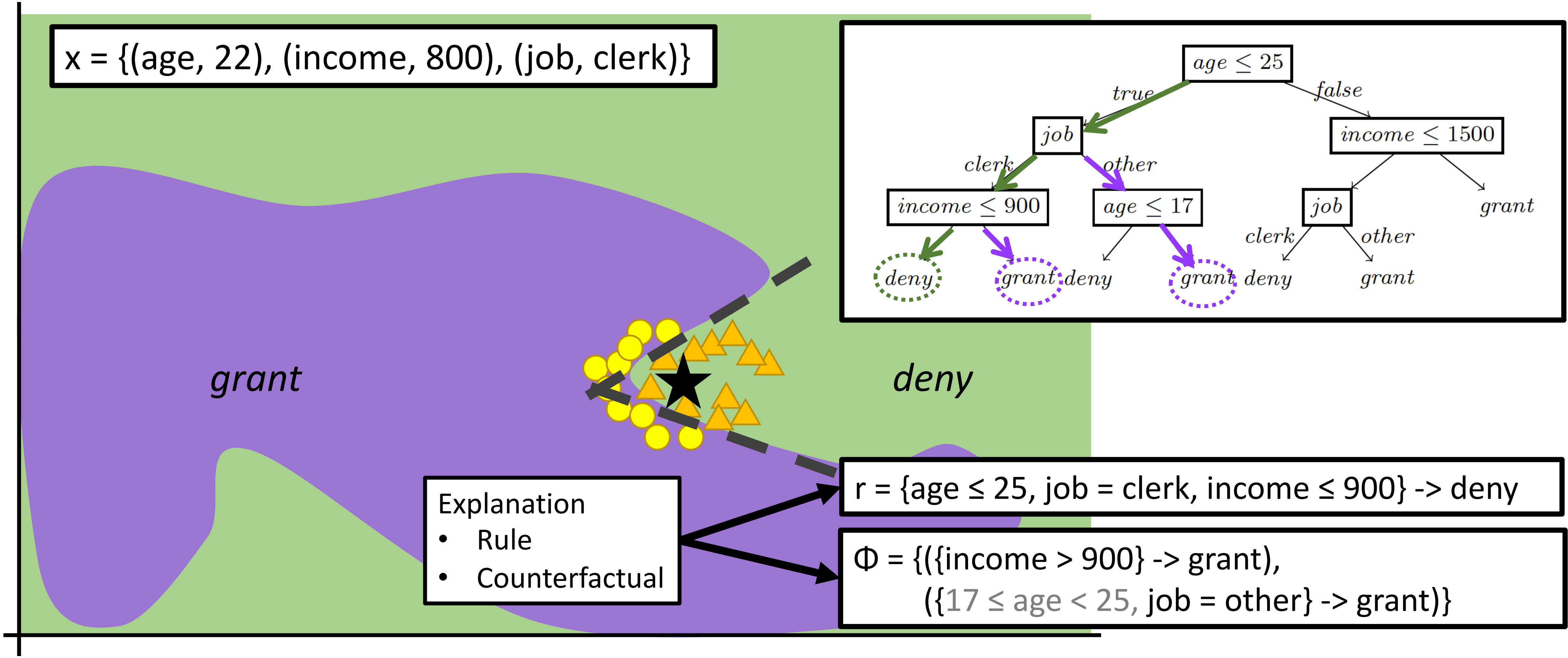
fitness

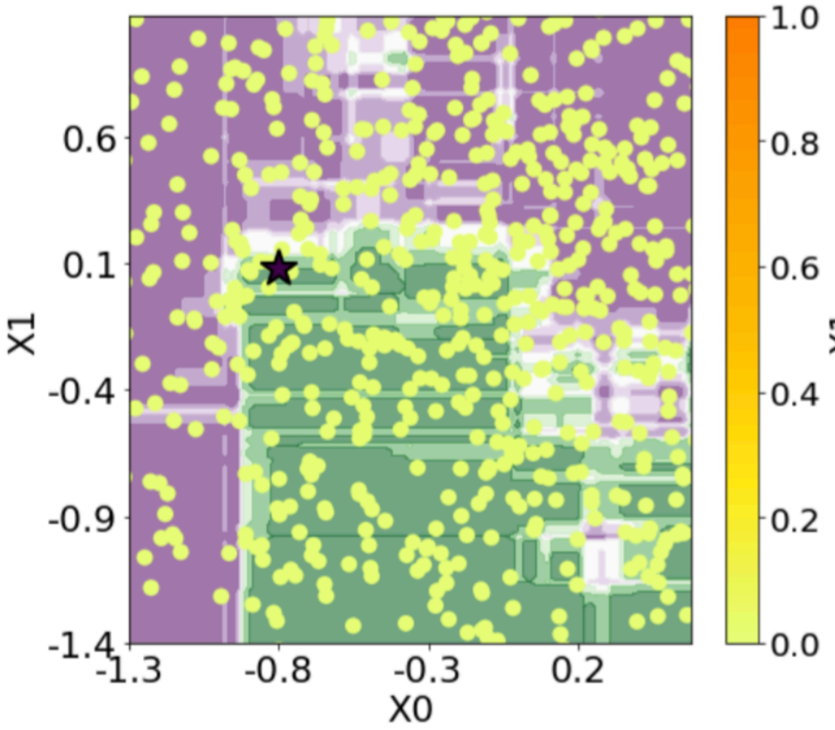
$$fitness^x_=(z) = I_{b(x)=b(z)} + (1 - d(x,z)) - I_{x=z}$$

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

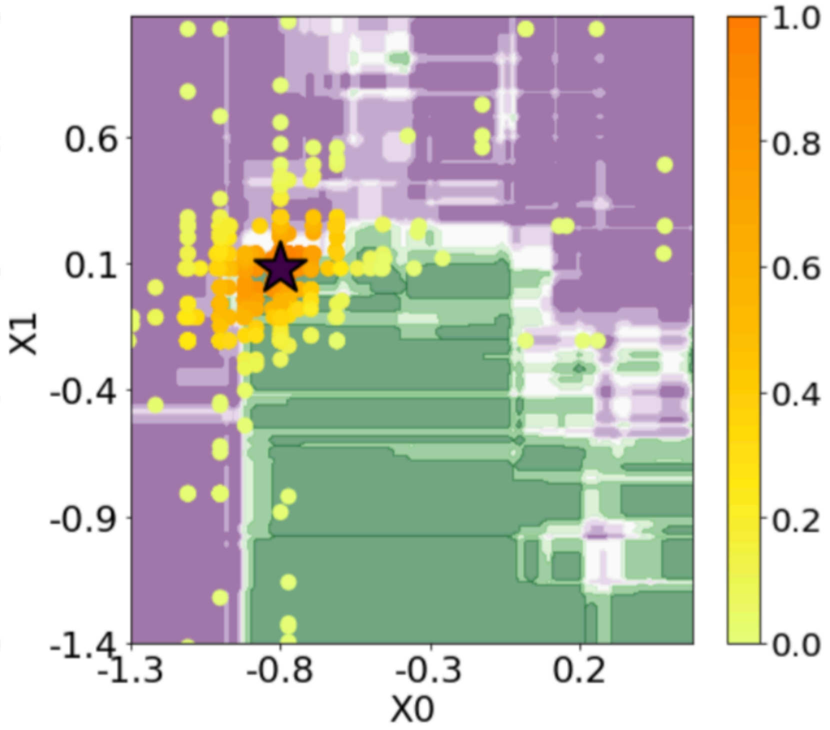
- Guidotti, R., Monreale, A., Ruggieri, S., Pedreschi, D., Turini, F., & Giannotti, F. (2018). ***Local Rule-Based Explanations of Black Box Decision Systems***. arXiv:1805.10820.

Local Rule-Based Explanations

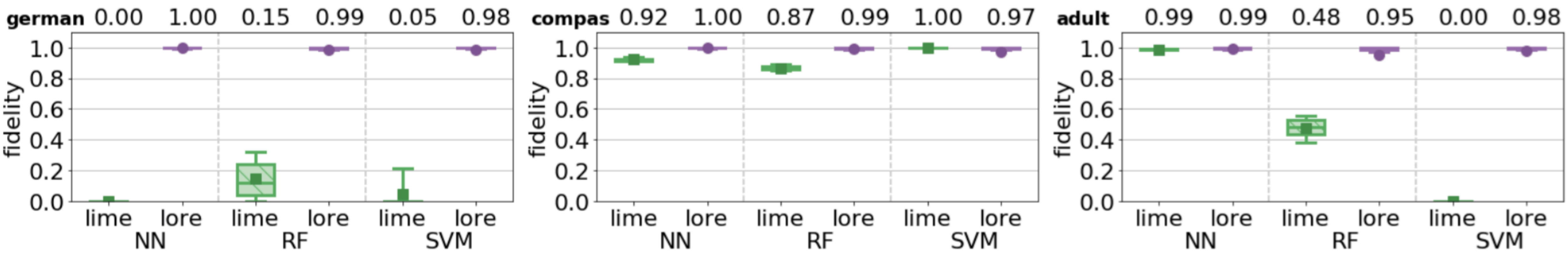




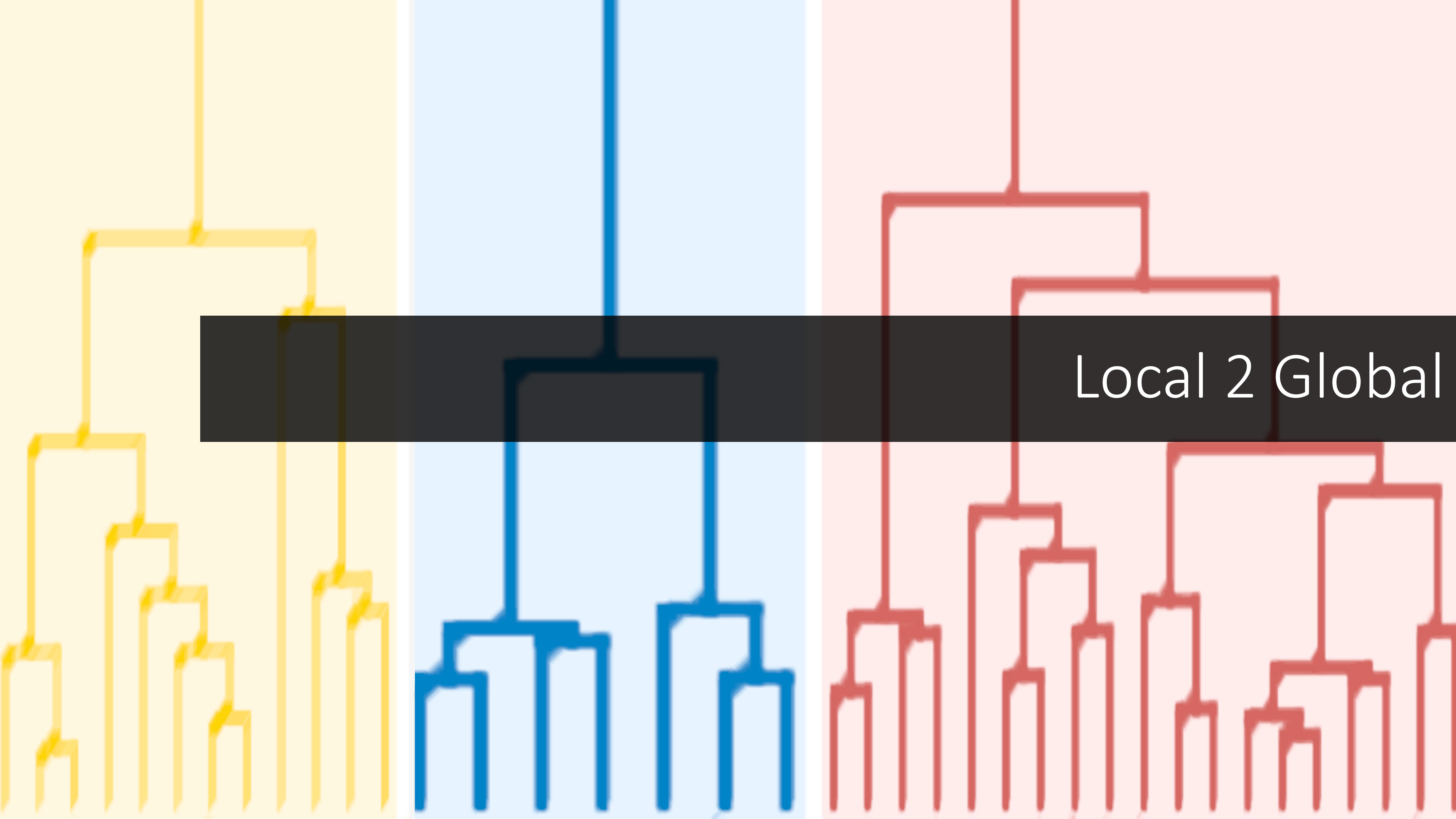
Random Neighborhood



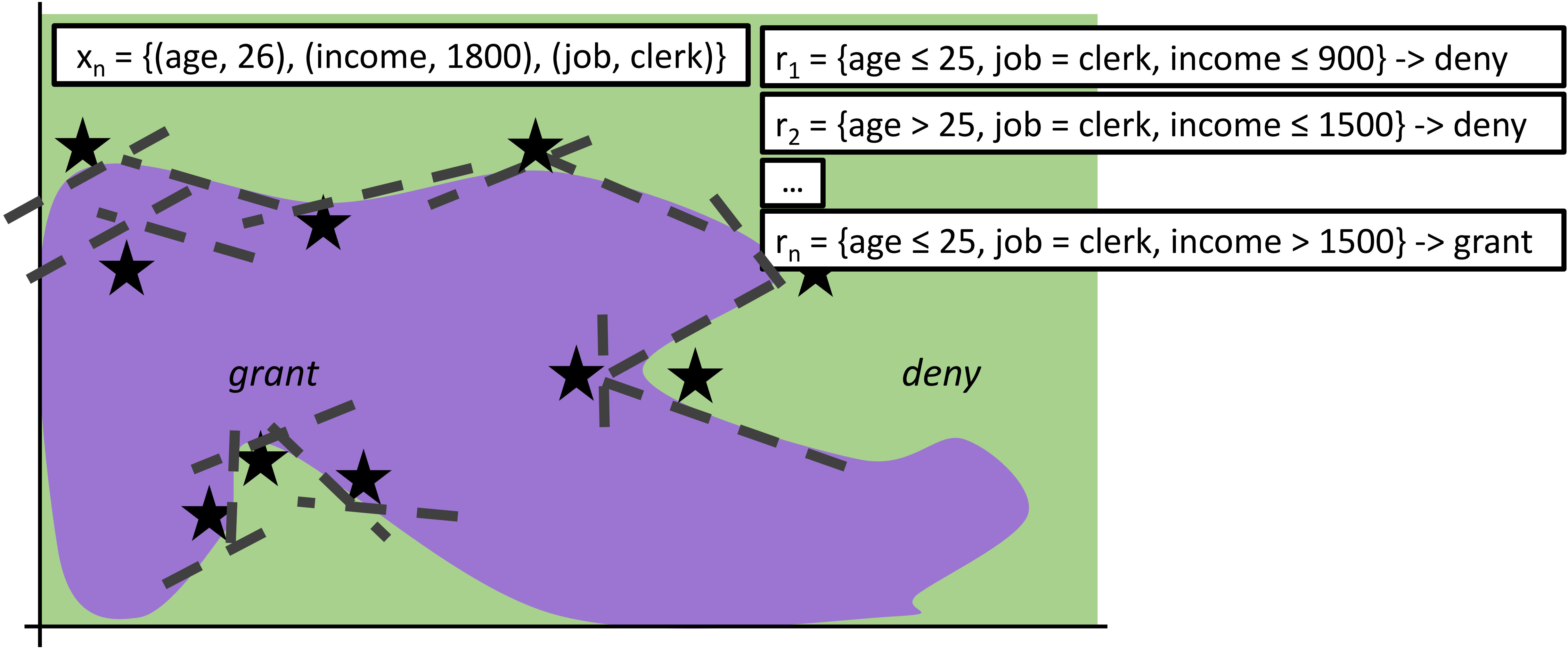
Genetic Neighborhood



Local 2 Global



Local First …



x_n = {(age, 26), (income, 1800), (job, clerk)}
r_1 = {age ≤ 25, job = clerk, income ≤ 900} -> deny
r_2 = {age > 25, job = clerk, income ≤ 1500} -> deny
…
r_n = {age ≤ 25, job = clerk, income > 1500} -> grant
grant
deny

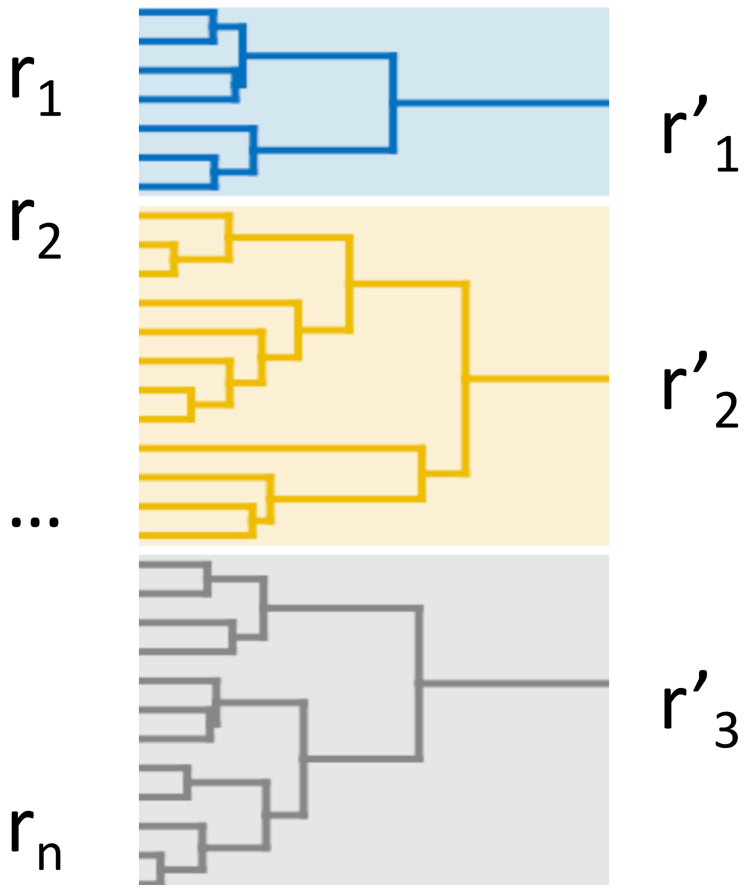
… then Local to Global

```
while score(fidelity, complexity) < α
    find similar theories
    merge them
```

Bayesian Information Criterion

Jaccard(coverage(T1), coverage(T2))

Union on concordant rules
Difference on discording rules



Meaningful Perturbations – SM, DNN, IMG

```
x instance to explain
varying x into x' maximizing b(x)~b(x')
the variation runs replacing a region R of x with:
     constant value, noise, blurred image
reformulation: find smallest R such that b(x_R)≪b(x)
```

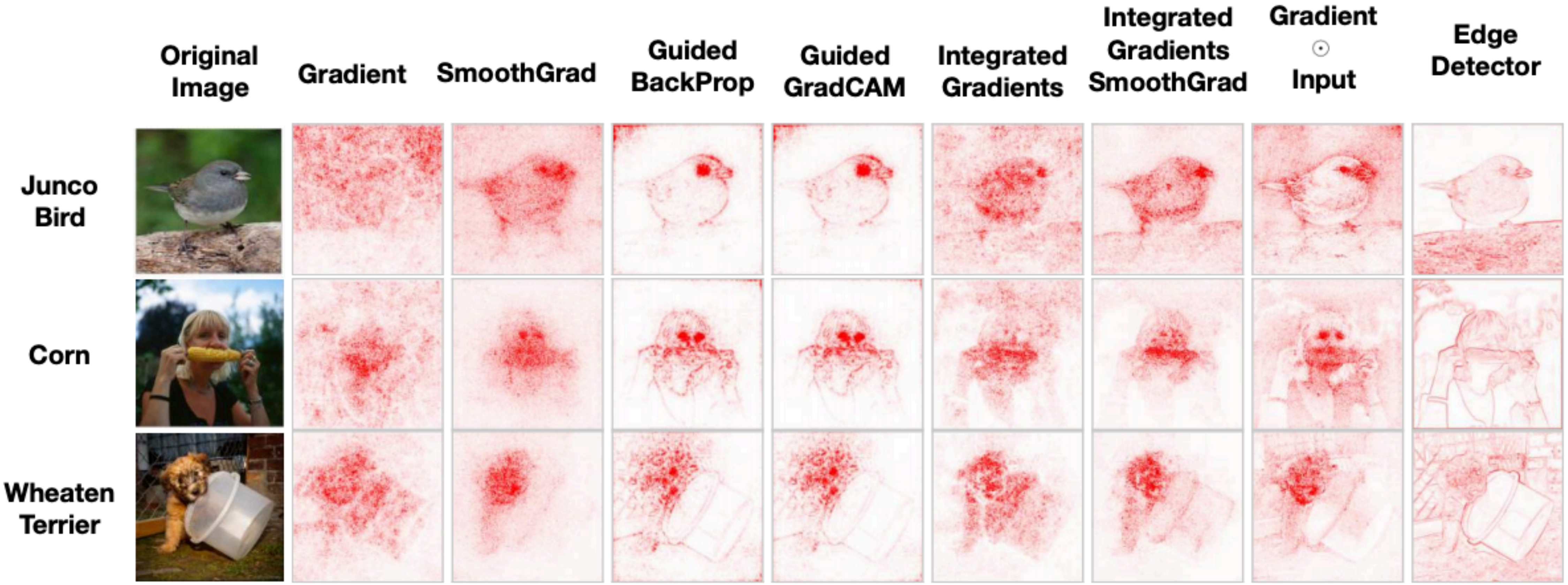
black box auditing

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

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			✓	✓	✓

Saliency maps



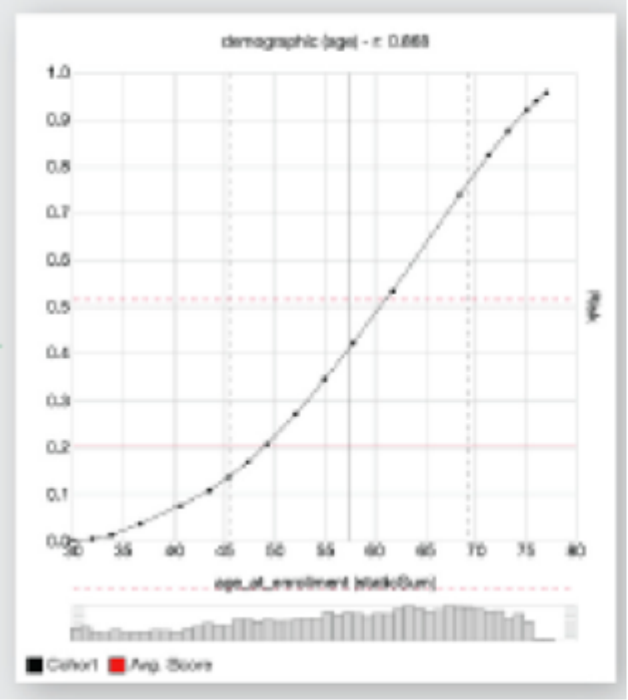
Julius Adebayo, Justin Gilmer, Michael Christoph Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. 2018.

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.

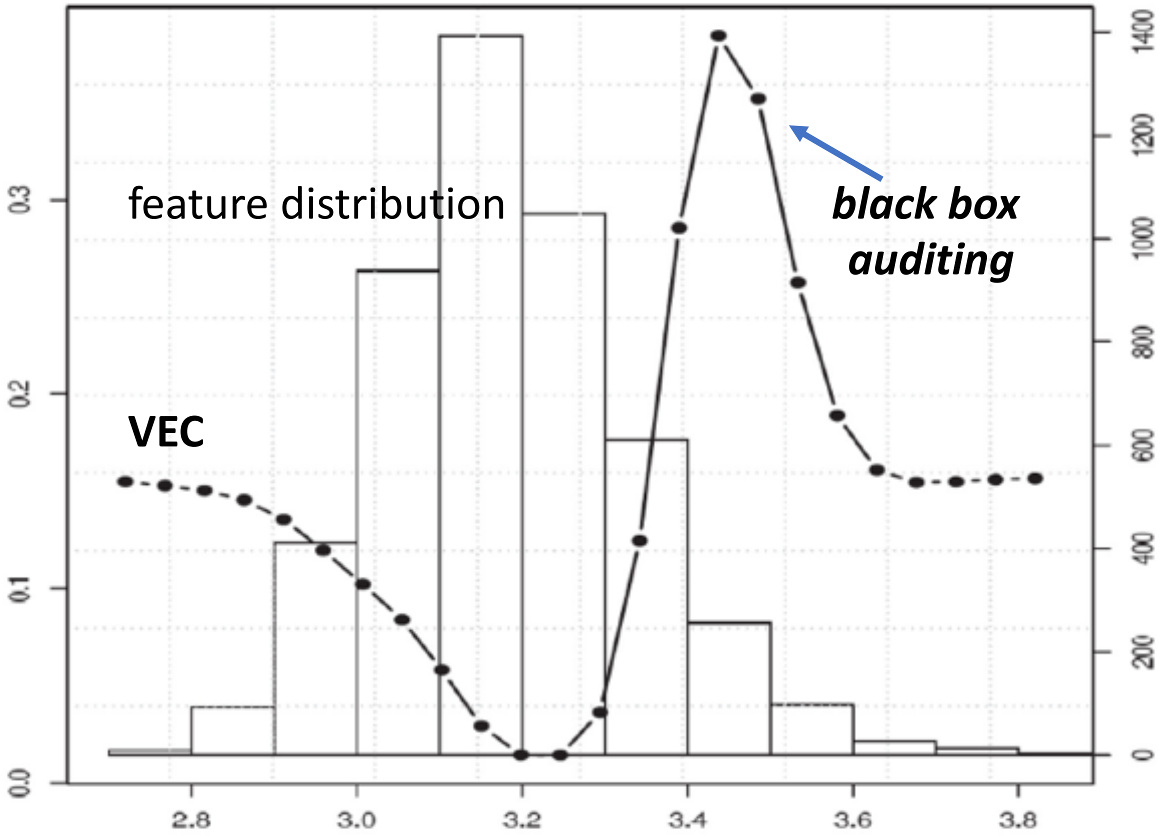
Inspection Model Explainers

- 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

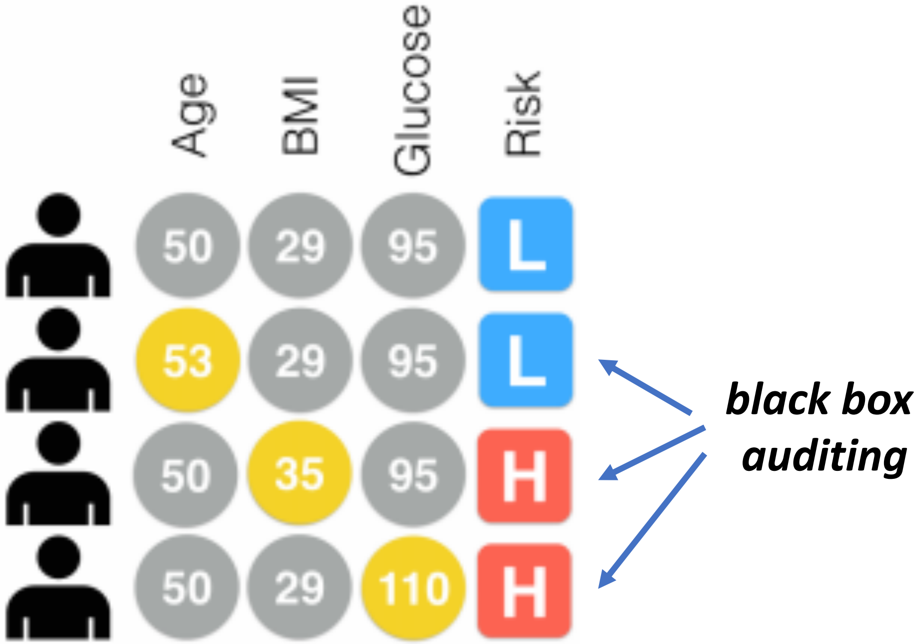
- Sensitivity measures are variables calculated as the range, gradient, variance of the prediction.
- The visualizations realized are barplots for the features importance, and ***Variable Effect Characteristic*** curve (VEC) plotting the input values versus the (average) outcome responses.

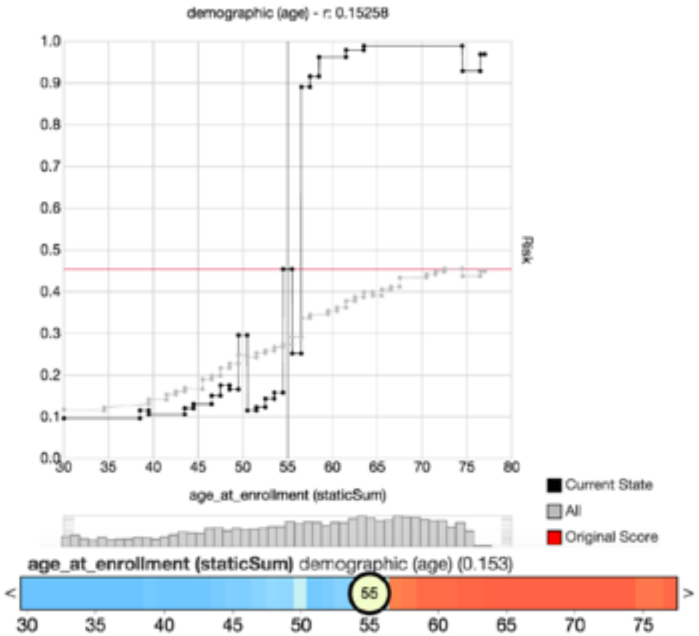


- Paulo Cortez and Mark J. Embrechts. 2011. ***Opening black box data mining models using sensitivity analysis***. CIDM.

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).

Name	Ref.	Authors	Year	Explanator	Black Box	Data Type	General	Random	Examples	Code	Dataset
CPAR	[135]	Yin et al.	2003	DR	–	TAB					✓
FRL	[127]	Wang et al.	2015	DR	–	TAB			✓	✓	✓
BRL	[66]	Letham et al.	2015	DR	–	TAB			✓		
TLBR	[114]	Su et al.	2015	DR	–	TAB			✓		✓
IDS	[61]	Lakkaraju et al.	2016	DR	–	TAB			✓		
Rule Set	[130]	Wang et al.	2016	DR	–	TAB			✓	✓	✓
1Rule	[75]	Malioutov et al.	2017	DR	–	TAB			✓		✓
PS	[9]	Bien et al.	2011	PS	–	ANY			✓		✓
BCM	[51]	Kim et al.	2014	PS	–	ANY			✓		✓
OT-SpAMs	[128]	Wang et al.	2015	DT	–	TAB			✓	✓	✓

Solving The Transparent Design Problem

Transparent Model Explainers

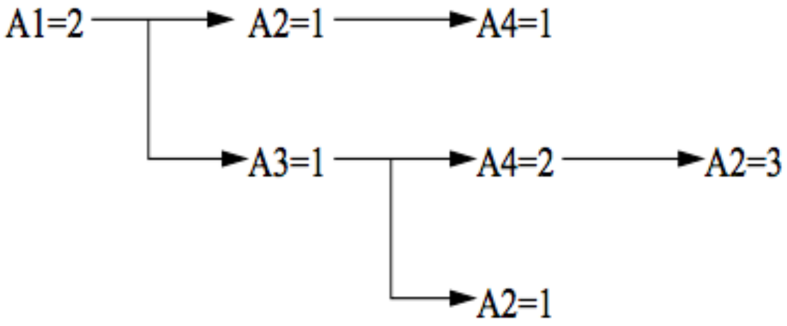
- Explanators:
  - DR
  - DT
  - PS

- Data Type:
  - TAB

CPAR – DR, TAB

- Combines the advantages of associative classification and rule-based classification.
- It adopts a greedy algorithm to generate ***rules directly from training data***.
- It generates more rules than traditional rule-based classifiers to ***avoid missing important rules***.
- To ***avoid overfitting*** it uses expected accuracy to evaluate each rule and uses the best *k* rules in prediction.

$$(A_1 = 2,\ A_2 = 1,\ A_4 = 1).$$
$$(A_1 = 2,\ A_3 = 1,\ A_4 = 2,\ A_2 = 3).$$
$$(A_1 = 2,\ A_3 = 1,\ A_2 = 1).$$



- Xiaoxin Yin and Jiawei Han. 2003. ***CPAR: Classification based on predictive association rules***. SIAM, 331–335

CORELS – DR, TAB

- It is a ***branch-and bound algorithm*** that provides the optimal solution according to the training objective with a certificate of optimality.
- It ***maintains a lower bound*** on the minimum value of error that each incomplete rule list can achieve. This allows to ***prune an incomplete rule list*** and every possible extension.
- It terminates with the optimal rule list and a certificate of optimality.

if $(age = 18 - 20)$ **and** $(sex = male)$ **then predict** yes
else if $(age = 21 - 23)$ **and** $(priors = 2 - 3)$ **then predict** yes
else if $(priors > 3)$ **then predict** yes
else predict no

- Angelino, E., Larus-Stone, N., Alabi, D., Seltzer, M., & Rudin, C. 2017. ***Learning certifiably optimal rule lists***. KDD.

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

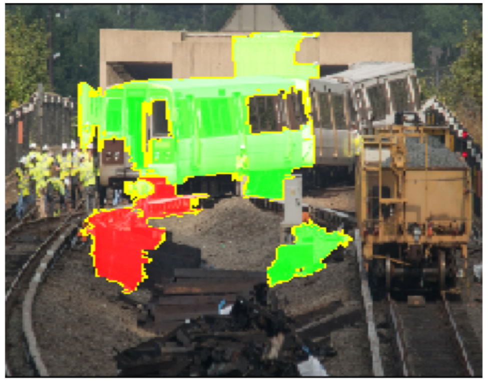
- M. Gethsiyal Augasta and T. Kathirvalavakumar. 2012. ***Reverse engineering the neural networks for rule extraction in classification problems***. NPL
- Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Dino Pedreschi, Franco Turini, and Fosca Giannotti. 2018. ***Local rule-based explanations of 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).

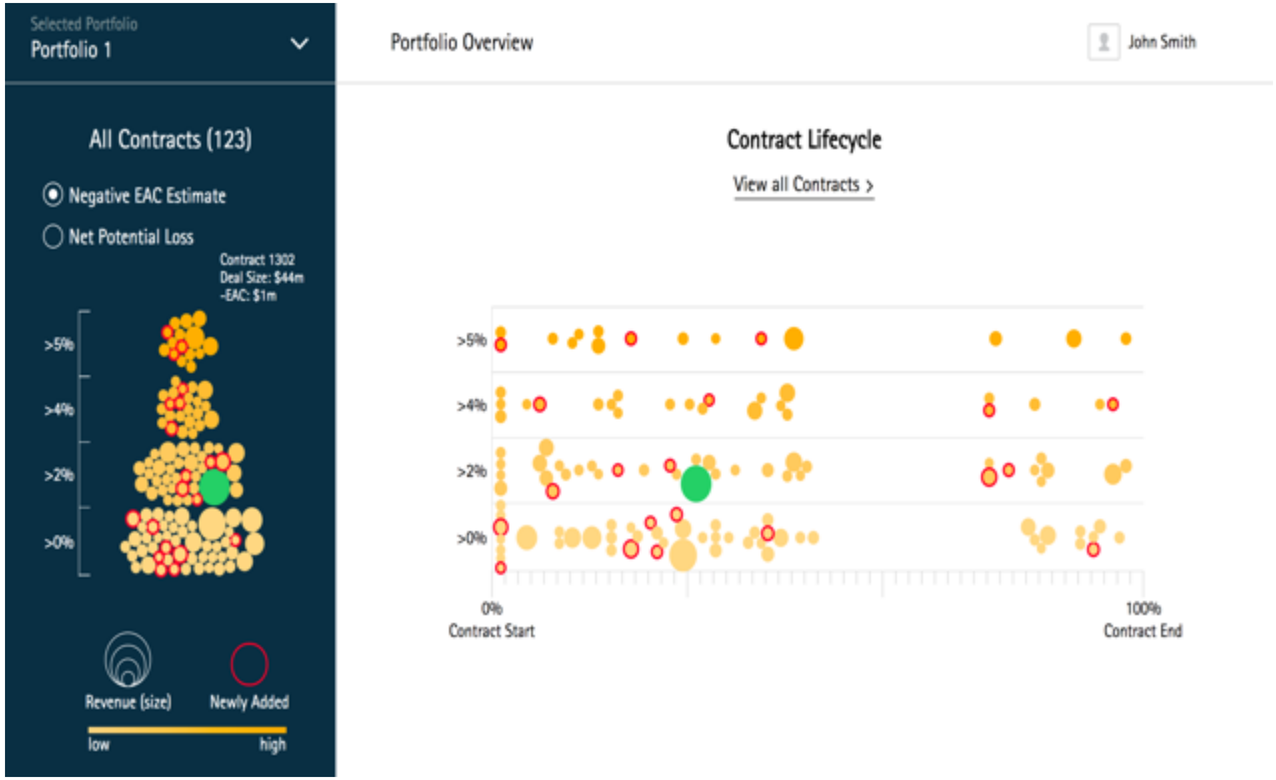
AI Technology: Integration of AI related technologies i.e., Machine Learning (Deep Learning / Recurrent neural Network), Reasoning (through semantics-augmented case-based reasoning) and Natural Language Processing for building a robust model which can (1) predict flight delays in minutes, (2) explain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



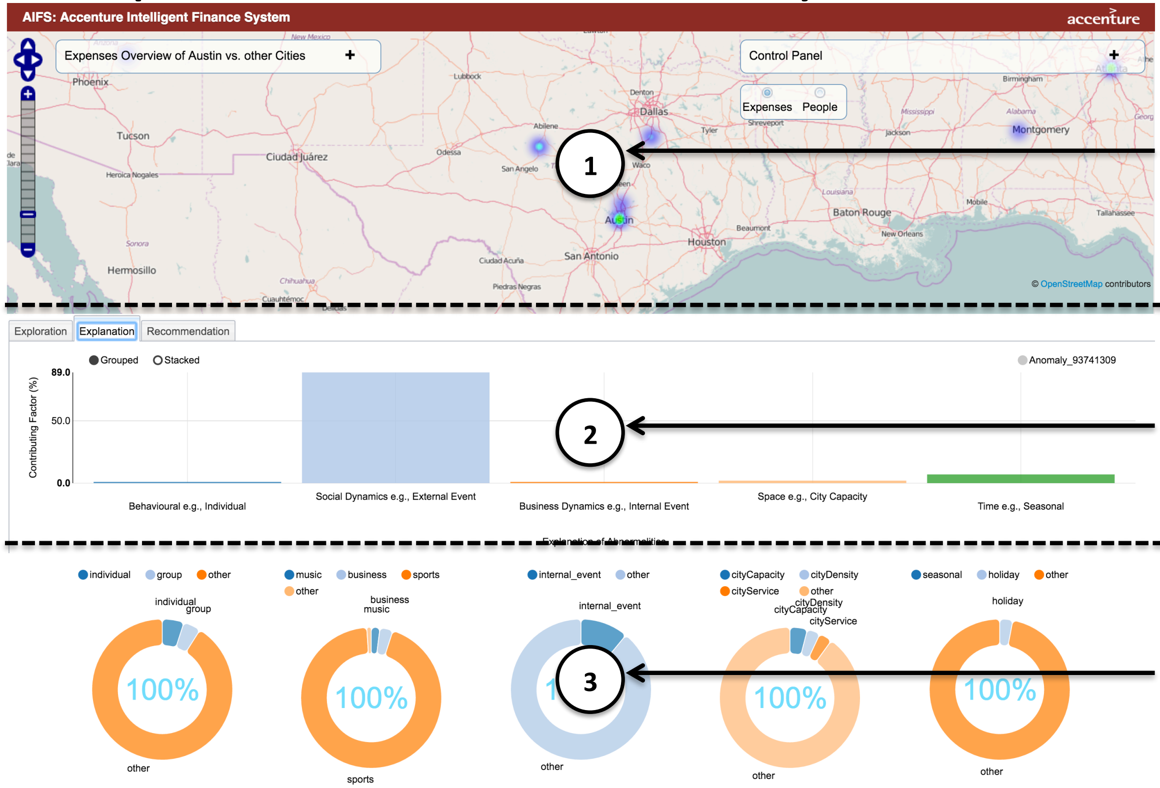
Jiewen Wu, Freddy Lécué, Christophe Guéret, Jer Hayes, Sara van de Moosdijk, Gemma Gallagher, Peter McCanney, Eugene Eichelberger: Personalizing Actions in Context for Risk Management Using Semantic Web Technologies. International Semantic Web Conference (2) 2017: 367-383

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

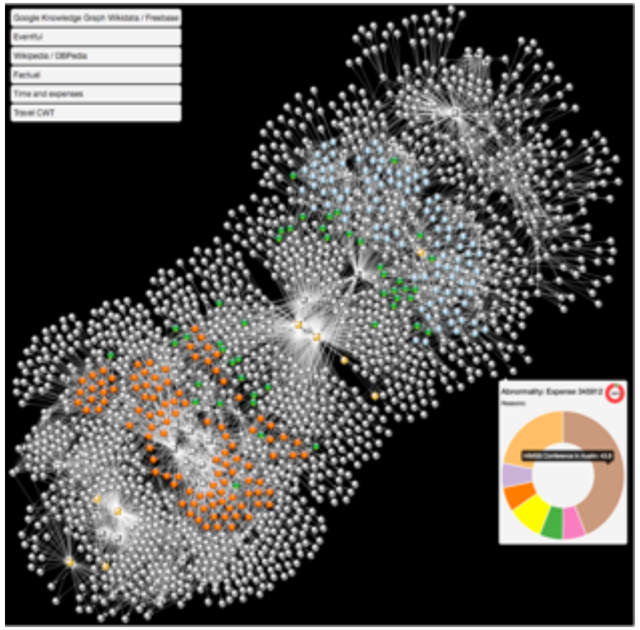
Explainable anomaly detection – Finance (Compliance)



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

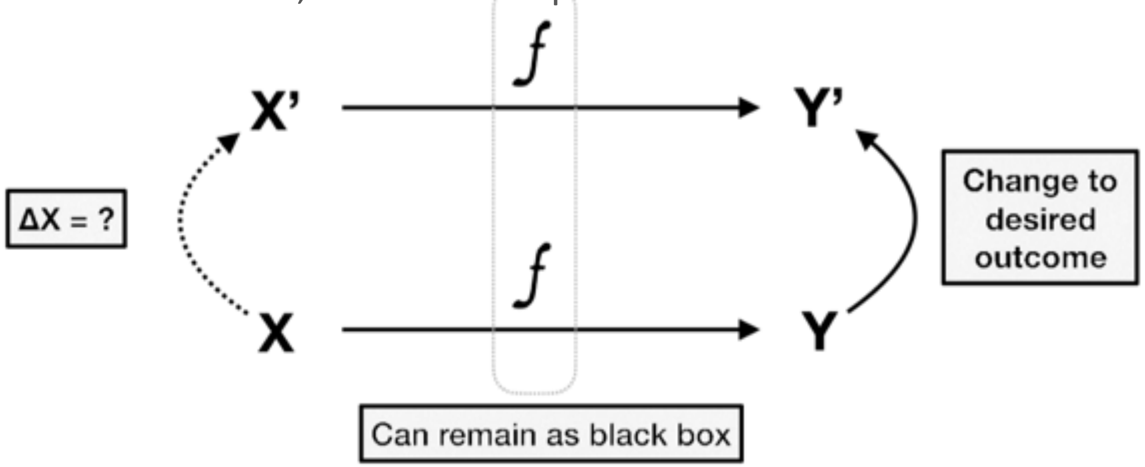
Counterfactual Explanations for Credit Decisions

- 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.

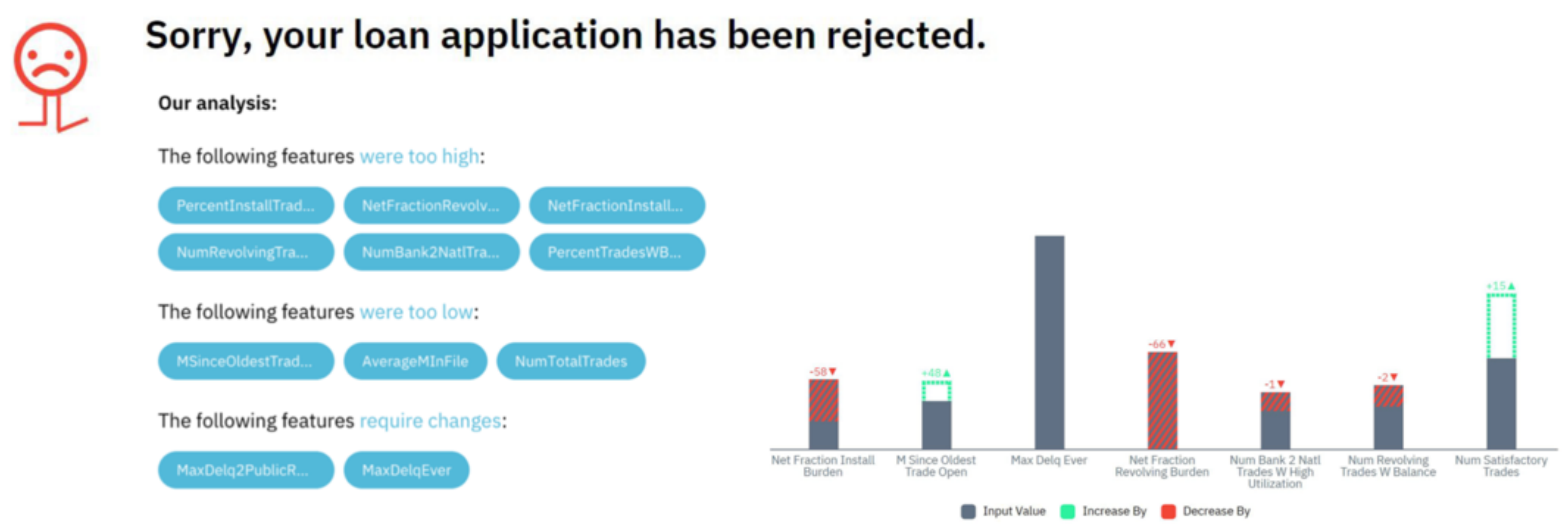
AI Technology: Supervised learning, binary classification.

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



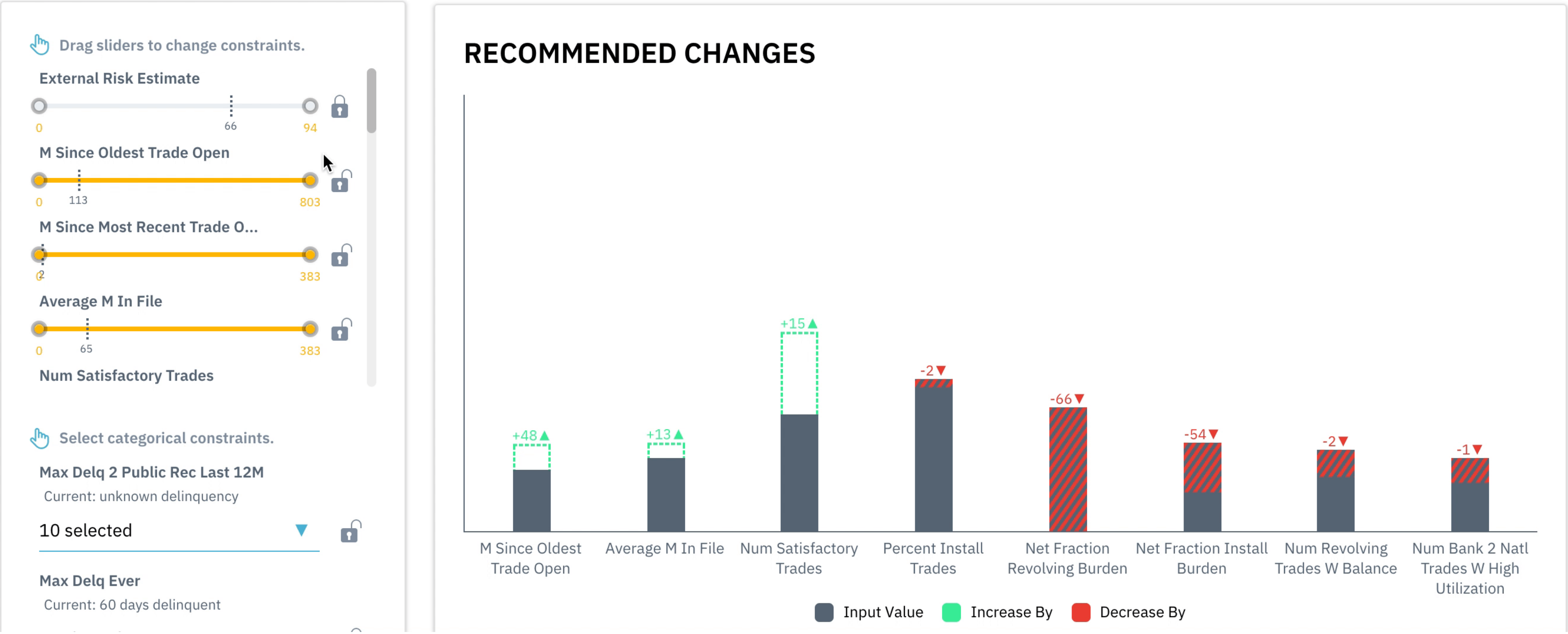
Rory Mc Grath, Luca Costabello, Chan Le Van, Paul Sweeney, Farbod Kamiab, Zhao Shen, Freddy Lécué: Interpretable Credit Application Predictions With Counterfactual Explanations. FEAP-AI4fin workshop, NeurIPS, 2018.

Counterfactual Explanations for Credit Decisions



Counterfactuals suggest where to increase (green, dashed) or decrease (red, striped) each feature.

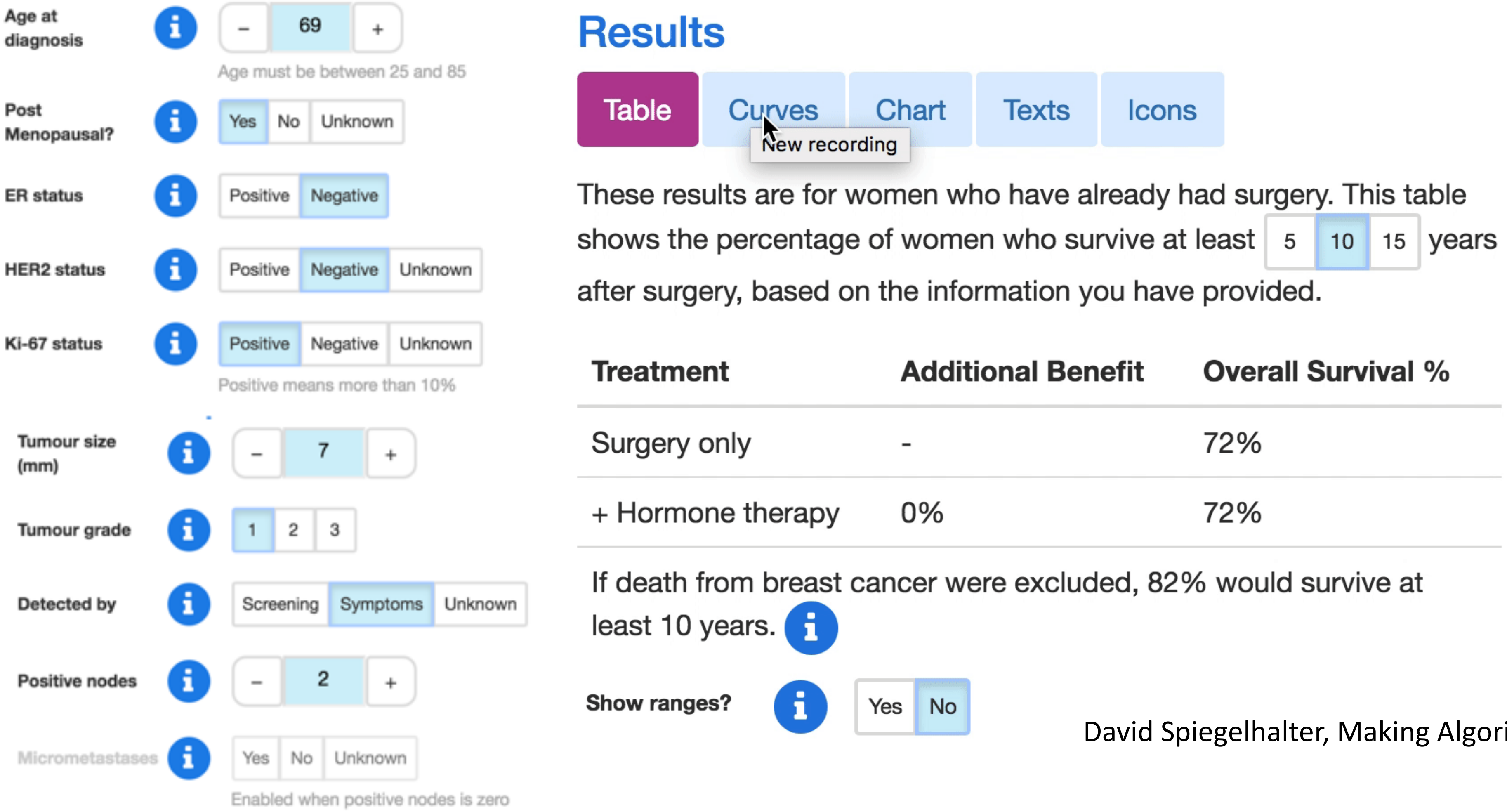
Rory Mc Grath, Luca Costabello, Chan Le Van, Paul Sweeney, Farbod Kamiab, Zhao Shen, Freddy Lécué: Interpretable Credit Application Predictions With Counterfactual Explanations. FEAP-AI4fin workshop, NeurIPS, 2018.



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

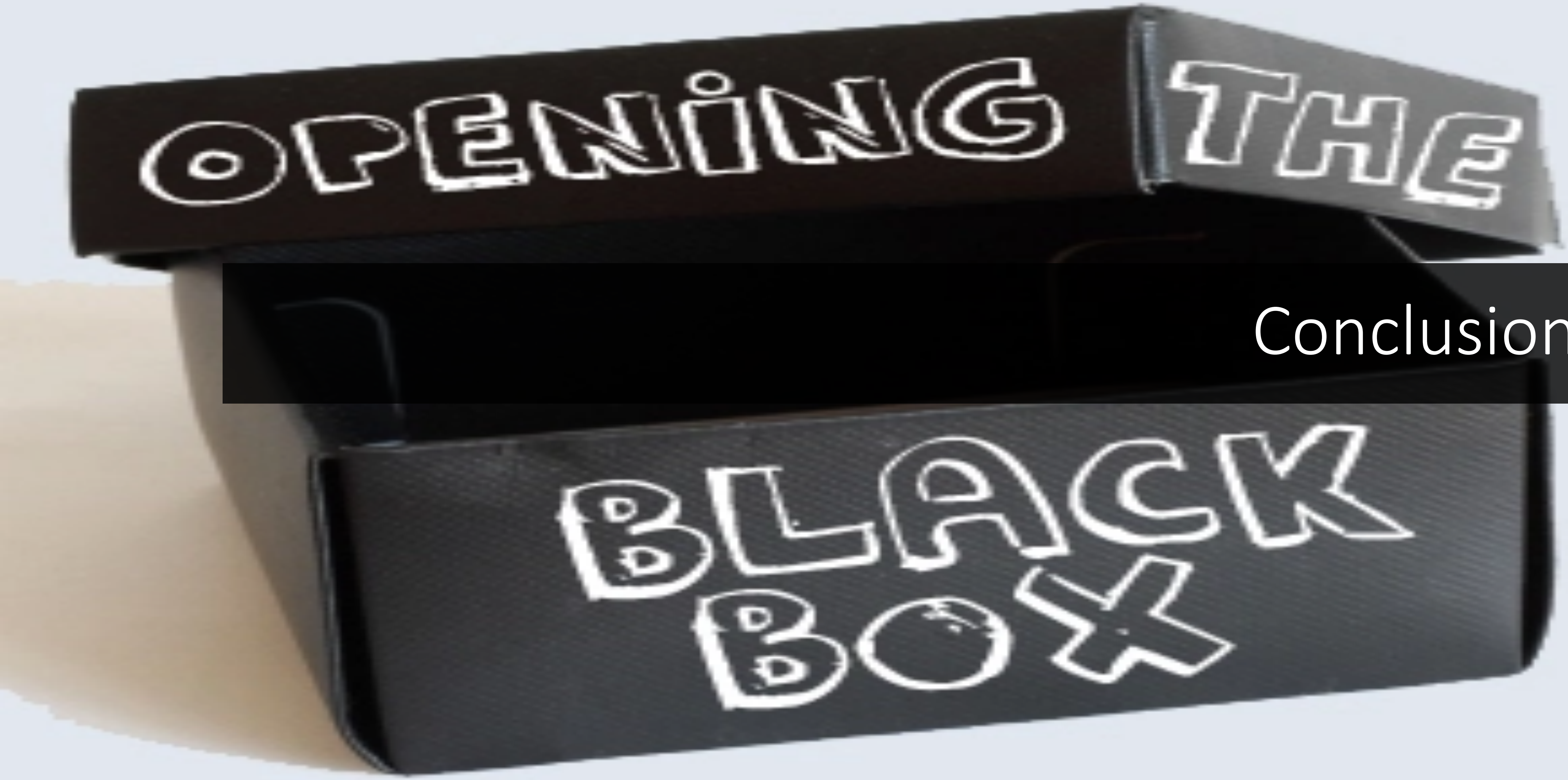
XAI Technology: Interactive explanations, Multiple representations.

David Spiegelhalter, Making Algorithms trustworthy, NeurIPS 2018 Keynote

predict.nhs.uk/tool

(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



Conclusions

Take-Home Messages

- Explainable AI is motivated by **real-world application of AI**
- Not a new problem – a reformulation of past research challenges in AI
- Multi-disciplinary: multiple AI fields, HCI, cognitive psychology, social science
- In Machine Learning:
  - Transparent design or post-hoc explanation?
  - Background knowledge matters!
- In AI (in general): many interesting / complementary approaches

Open The Black Box!

- ***To empower*** individual against undesired effects of automated decision making
- ***To implement*** the “right of explanation”
- ***To improve*** industrial standards for developing AI-powered products, increasing the trust of companies and consumers
- ***To help*** people make better decisions
- ***To preserve*** (and expand) human autonomy

Open Research Questions

- There is ***no agreement*** on ***what an explanation is***
- There is ***not a formalism*** for ***explanations***
- There is ***no work*** that seriously addresses the problem of ***quantifying*** the grade of ***comprehensibility*** of an explanation for humans
- What happens when black box make decision in presence of ***latent features***?
- What if there is a ***cost*** for querying a black box?

Future Challenges

- Creating awareness! Success stories!
- Foster multi-disciplinary collaborations in XAI research.
- Help shaping industry standards, legislation.
- More work on transparent design.
- Investigate symbolic and sub-symbolic reasoning.

- *Evaluation:*
    - *We need benchmark* - Shall we start a task force?
    - *We need an XAI challenge* - Anyone interested?
    - *Rigorous, agreed upon, human-based* evaluation protocols

Explainable AI:

From Theory to Motivation, Applications and Limitations

We hire!! Postdocs wanted

AI4EU

SoBigData

HUMANE AI

XAI

European Research Council
Established by the European Commission

http://ai4eu.org/

http://www.sobigdata.eu/

http://www.humane-ai.eu/

ERC-AdG-2019 "Science & technology for the eXplanation of AI decision making"