https://kdd.isti.cnr.it/xkdd2019/

Tutorial on

eXplainable Knowledge Discovery in Data Mining

Anna Monreale, Riccardo Guidotti, Pasquale Minervini, Salvo Rinzivillo

ISTITUTO DI SCIENZA E TECNOLOGIE DELL'INFORMAZIONE "A. FAEDO"

UNIVERSITÀ DI PISA

Tutorial Outline

- 10:30 - **Welcome and General Overview** - Anna Monreale
- 10:35 - **Science & Technology for AI Decision Making** - Anna Monreale
- 10:40 - **Explaining Explanation Methods** - Riccardo Guidotti
- 11:20 - **Explaining with Knowledge Graphs** - Pasquale Minervini
- 11:50 - **Explaining Privacy Risks** - Anna Monreale
- 12:20 - **Visualizing Explanations** - Riccardo Guidotti (Salvo Rinzivillo)
- 12:30 - **Conclusions and Q&A** - Anna Monreale
- 12:40 - Lunch break
- 14:00 - Workshop

Science & Technology for AI Decision Making

Anna Monreale, University of Pisa, Pisa

Definitions

Oxford Dictionary of English

explanation | ɛksplə'neɪʃ(ə)n |

noun

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

interpret | ɪn'tə:prɪt |

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

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

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**.
- **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

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

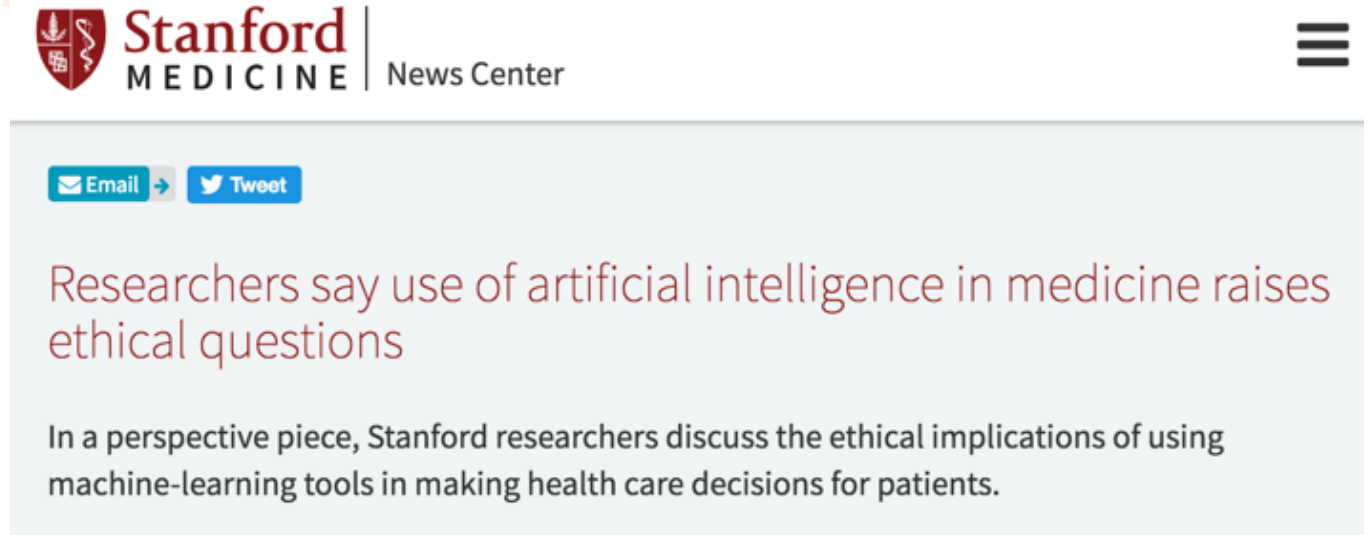
OP-ED CONTRIBUTOR

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

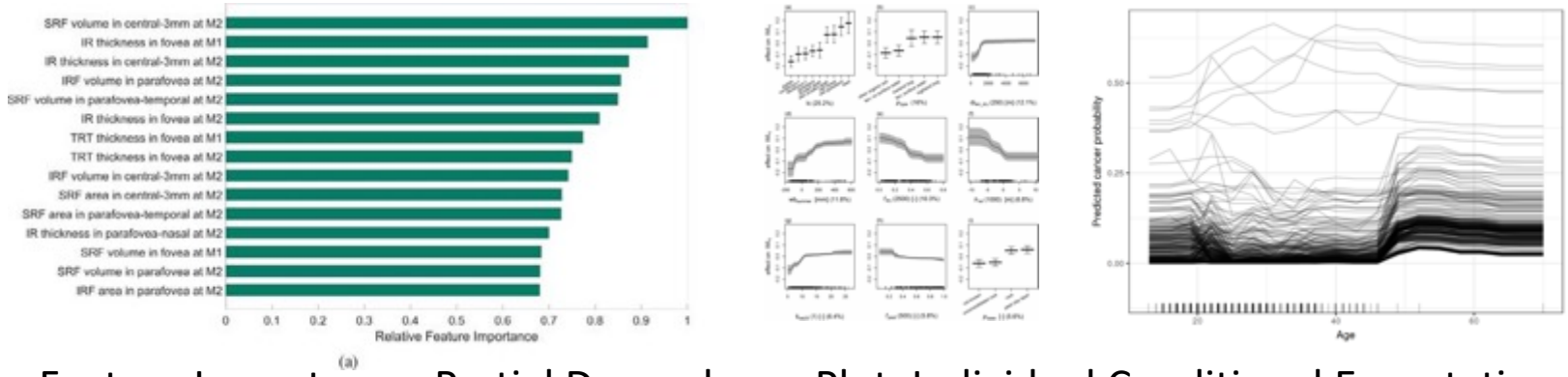
In a perspective piece, Stanford researchers discuss the ethical implications of using machine-learning tools in making health care decisions for patients.

Right of Explanation

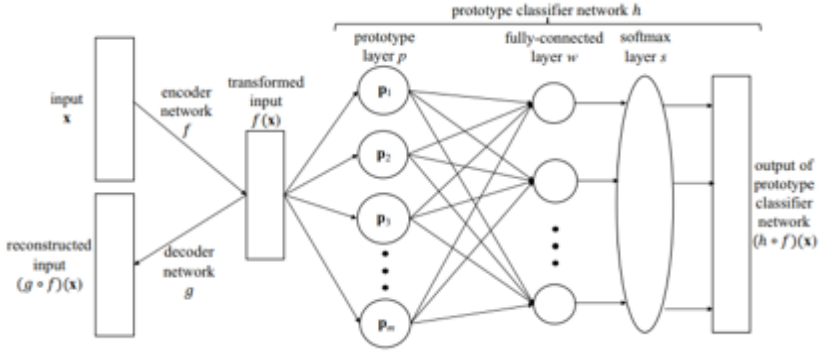
Since 25 May 2018, GDPR establishes a right for all individuals to obtain “meaningful explanations of the logic involved” when “automated (algorithmic) individual decision-making”, including profiling, takes place.

Explanation in different AI fields

- Machine Learning

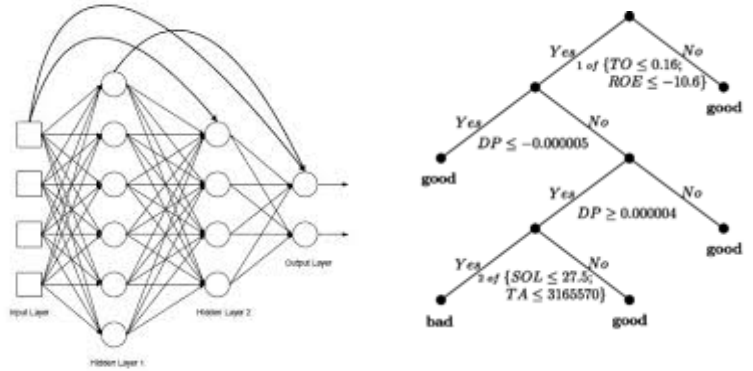


Feature Importance, Partial Dependence Plot, Individual Conditional Expectation



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

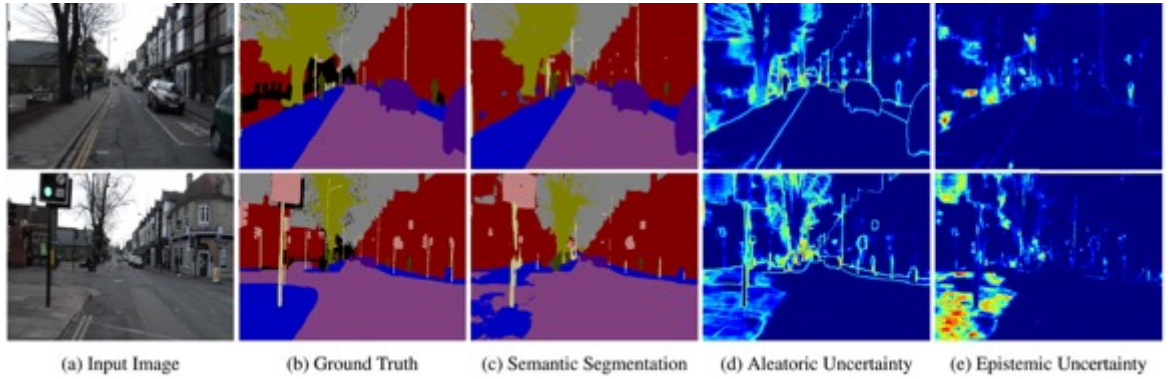


Surogate Model

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

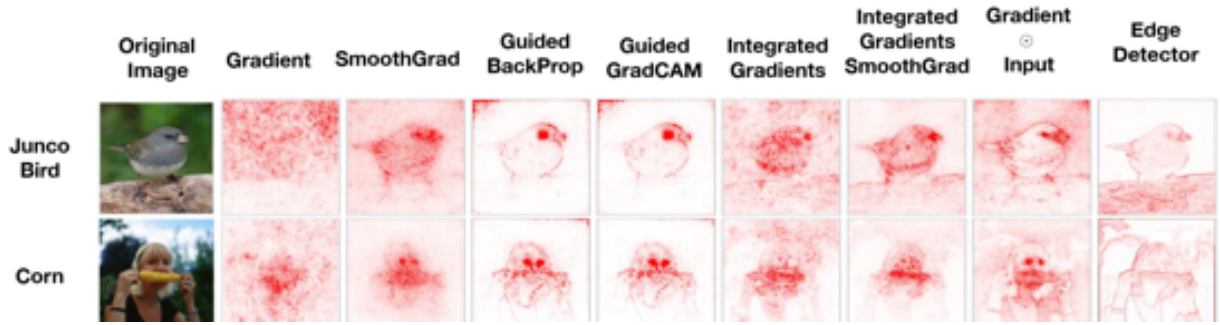
Explanation in different AI fields

- 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

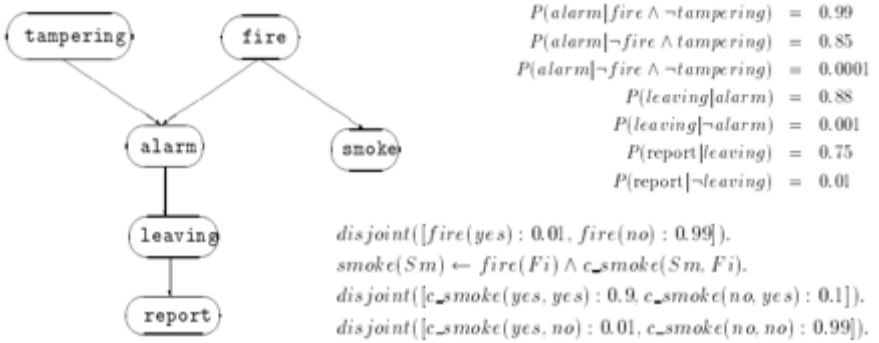


Saliency Map

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

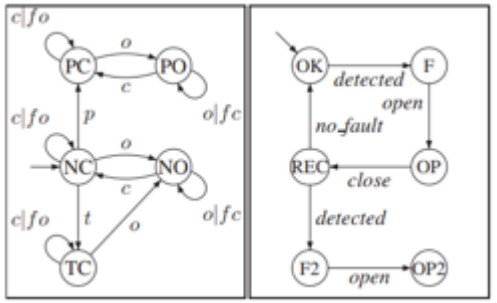
Explanation in different AI fields

- Machine Learning
- Computer Vision
- Knowledge Representation and Reasoning



Abduction Reasoning (in Bayesian Network)

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

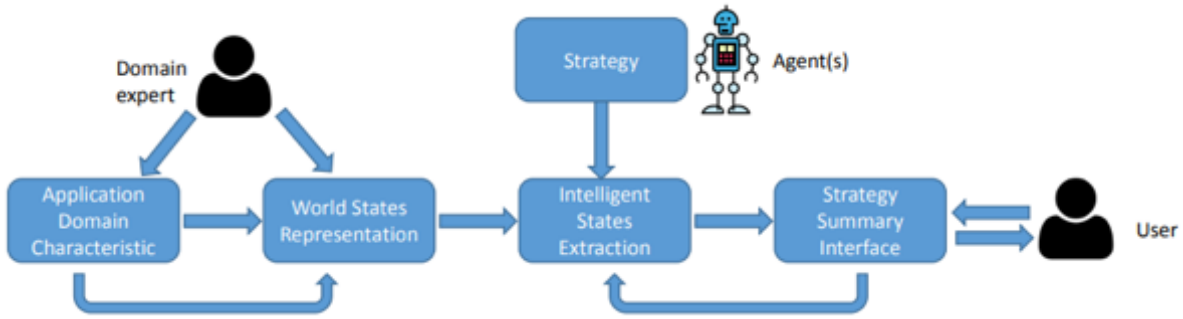


Diagnosis Inference

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

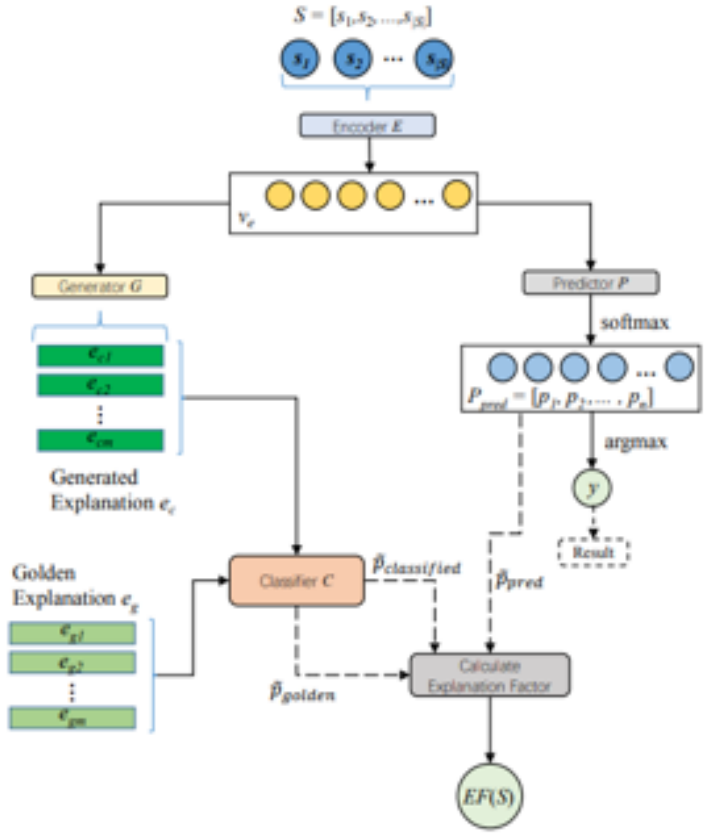
Ofra Amir, Finale Doshi-Velez, David Sarne: Agent Strategy Summarization. AAMAS 2018: 1203-1207

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

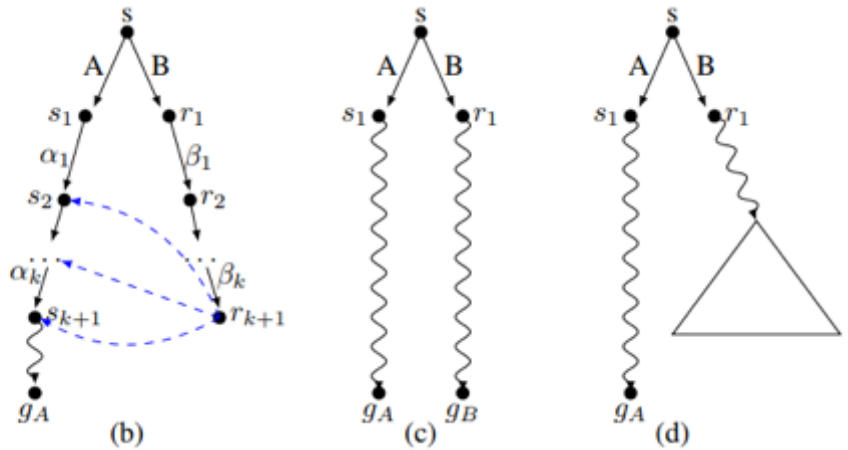


Explainable NLP

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?

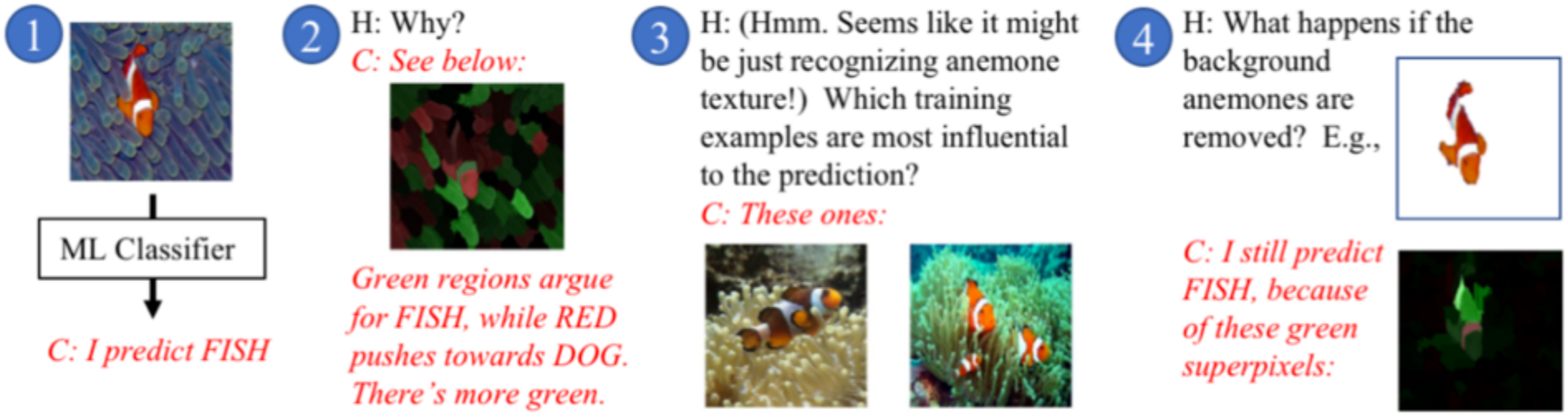
Robot: Here is the histogram of all training examples that reached this leaf. 80% of examples where this area was above 20 cm predicted the appropriate action to be "drive forward".

From Decision Tree to human-friendly information

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

Explanation as *Machine-Human Conversation*

[Weld and Bansal 2018]



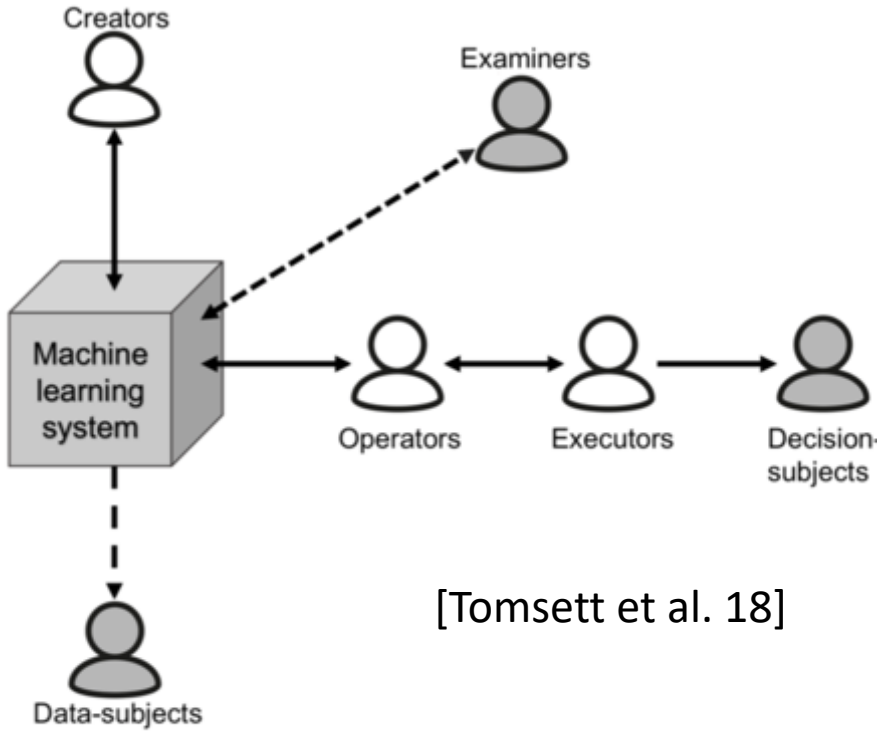
- Humans may have follow-up questions
- Explanations cannot answer all users' concerns

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]

Summarizing: the Need to Explain comes from …

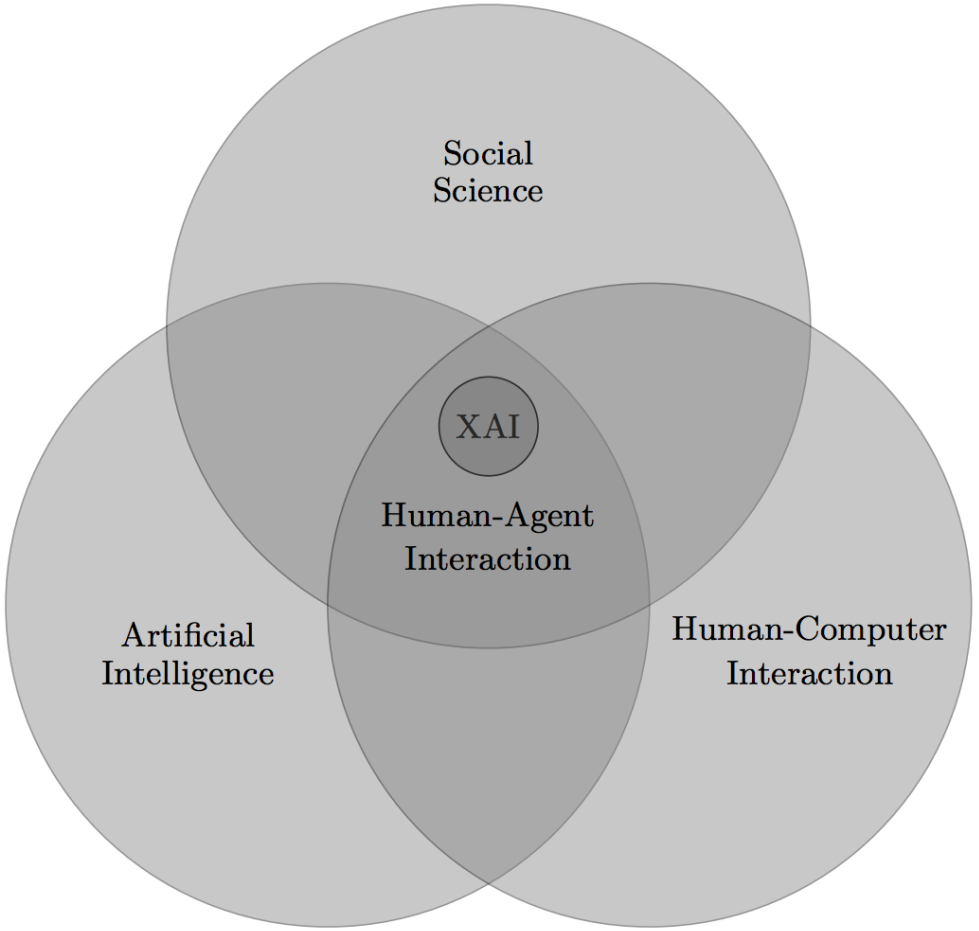
- User Acceptance & Trust [Lipton 2016, Ribeiro 2016, Weld and Bansal 2018]

- 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

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).

References

- [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).

Explaining Explanation Methods

Riccardo Guidotti, ISTI-CNR, Pisa

What is a Black Box Model?

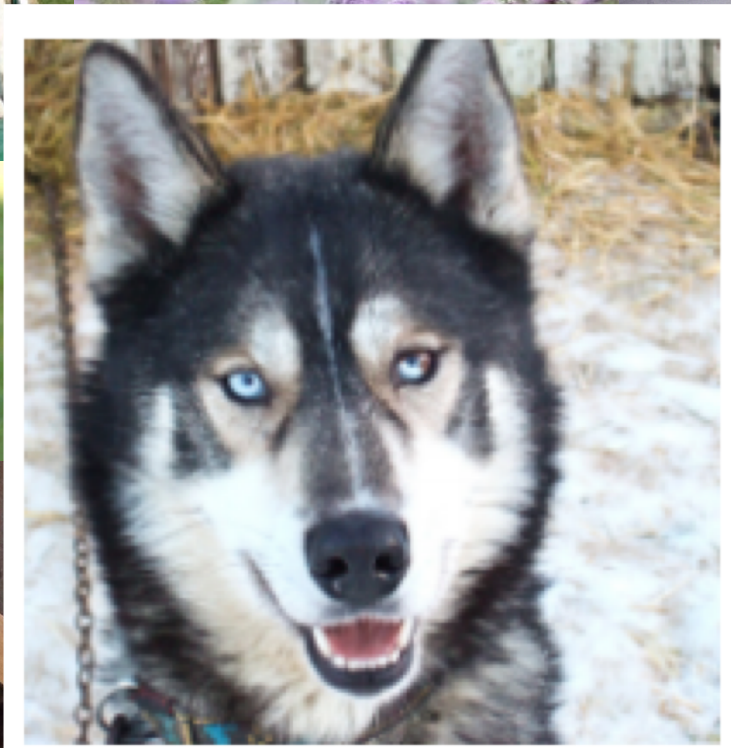
A ***black box*** is a model, whose internals are either unknown to the observer or they are known but uninterpretable by humans.

- 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.

Needs For Interpretable Models

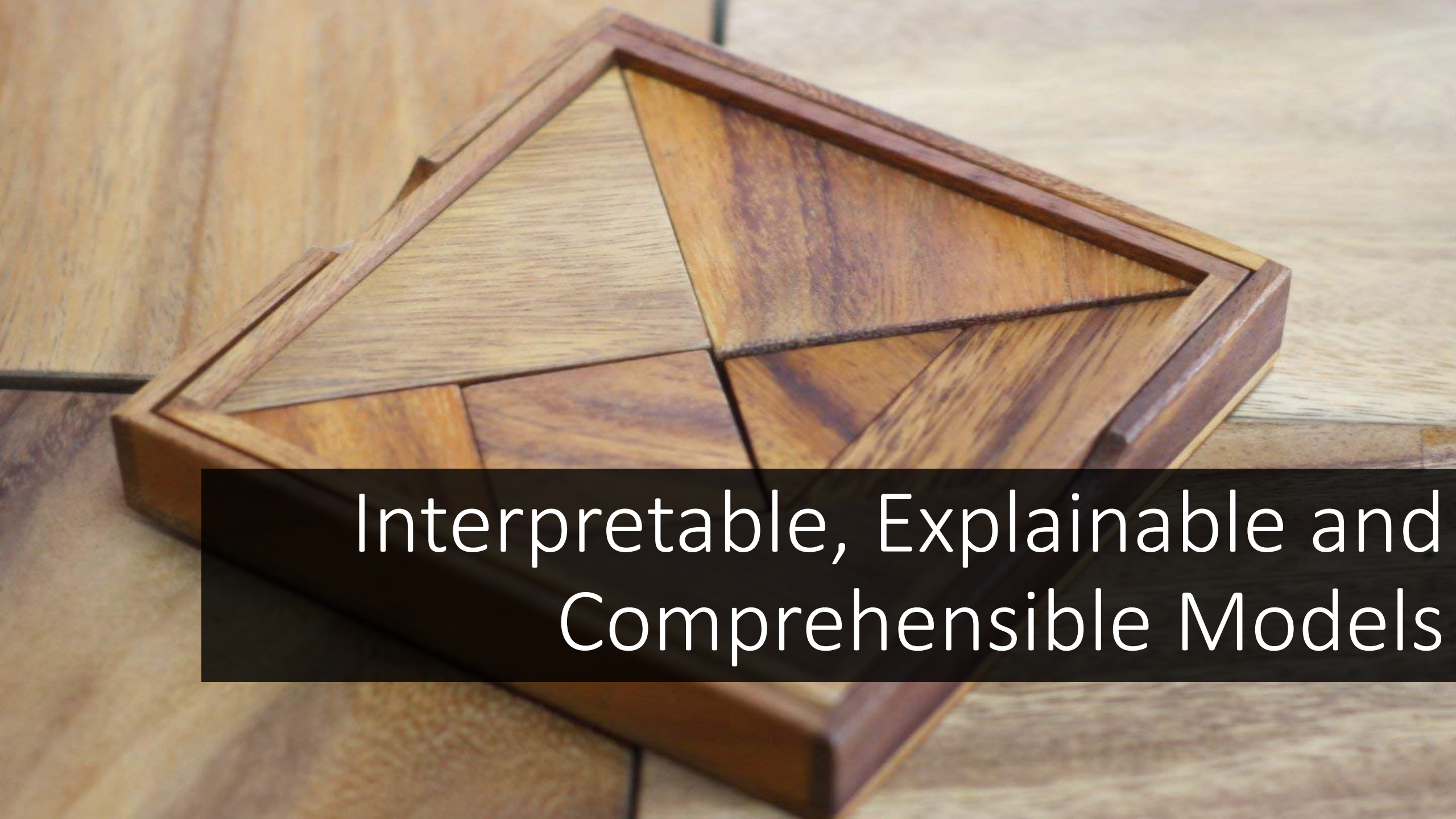
COMPAS recidivism black bias

The background bias



(a) Husky classified as wolf

(b) Explanation



Interpretable, Explainable and Comprehensible 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/

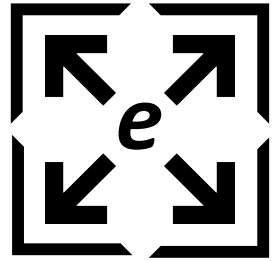
- Finale Doshi-Velez and Been Kim. 2017. ***Towards a rigorous science of interpretable machine learning***. arXiv:1702.08608v2.

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

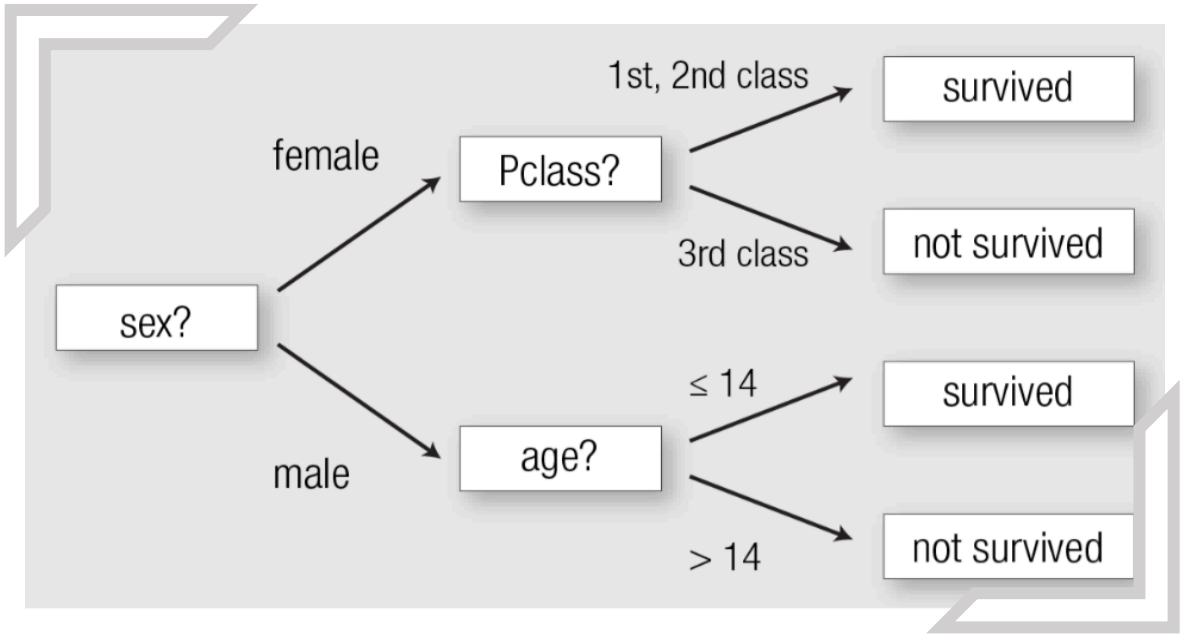
- ***Fairness***: the model guarantees the protection of groups against discrimination.
- ***Privacy***: the model does not reveal sensitive information about people.
- ***Respect Monotonicity***: the increase of the values of an attribute either increase or decrease in a monotonic way the probability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

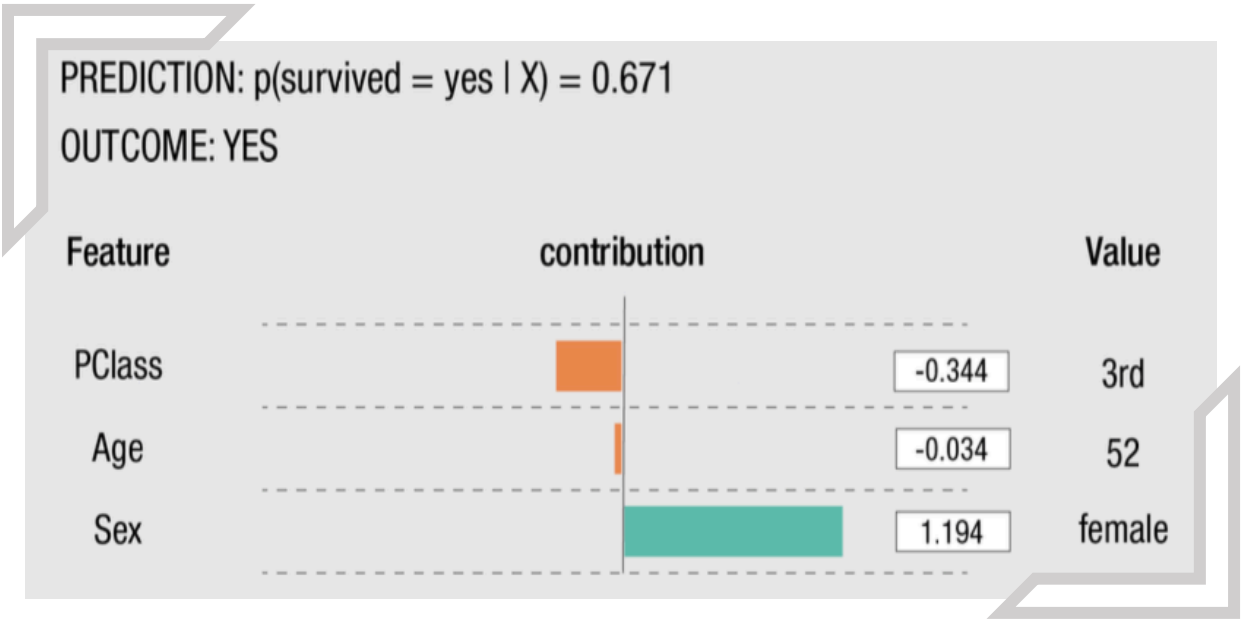
Desiderata of an Interpretable Model

- ***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

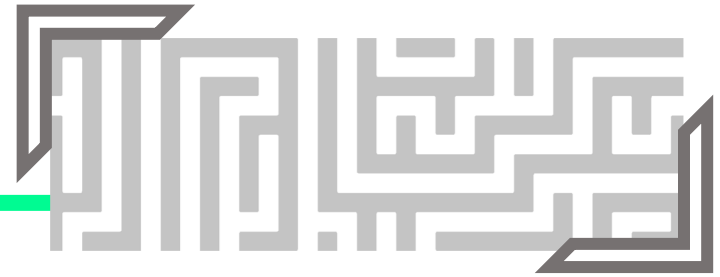


Linear Model

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

Rules

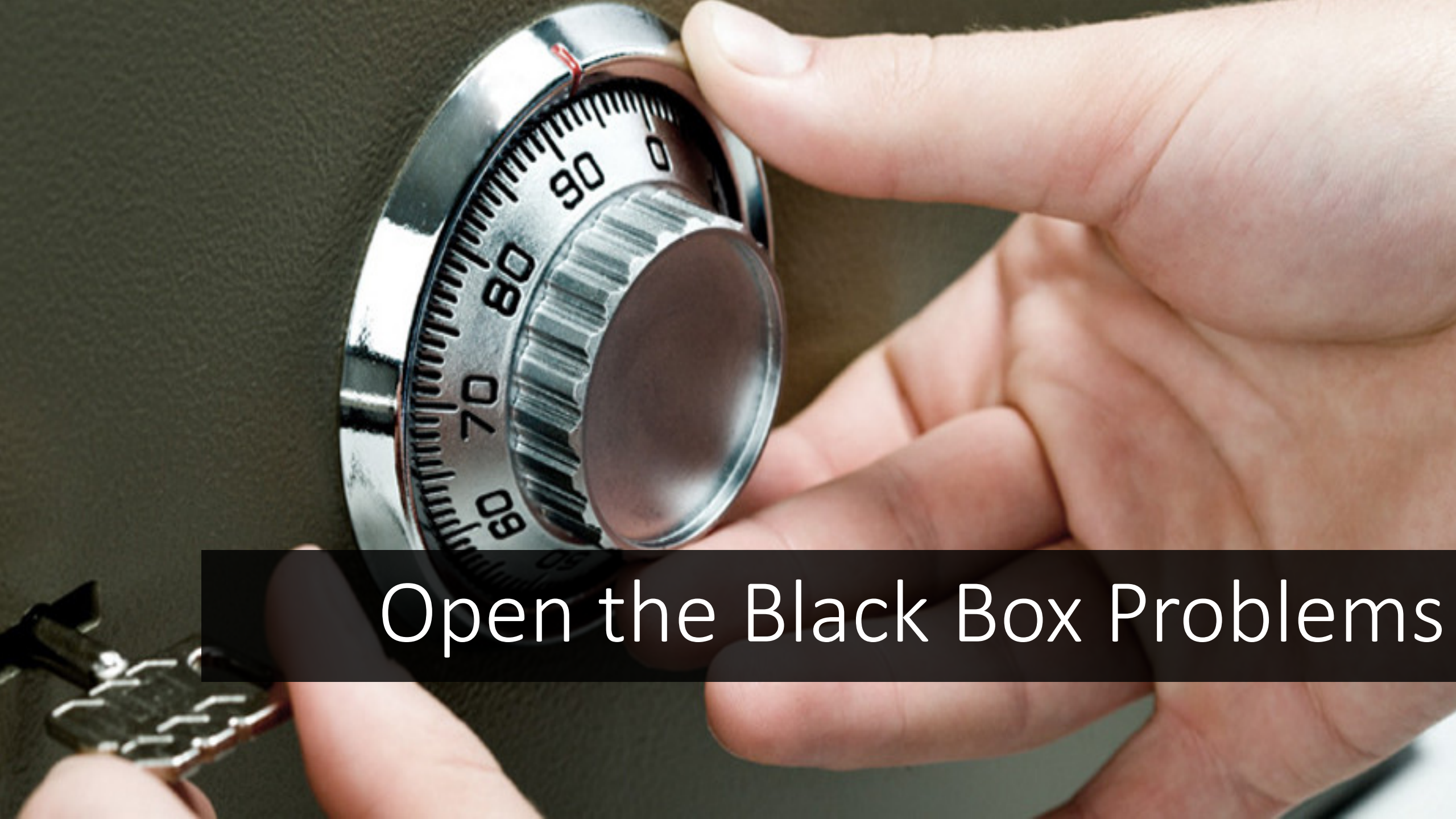
Complexity



- 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.

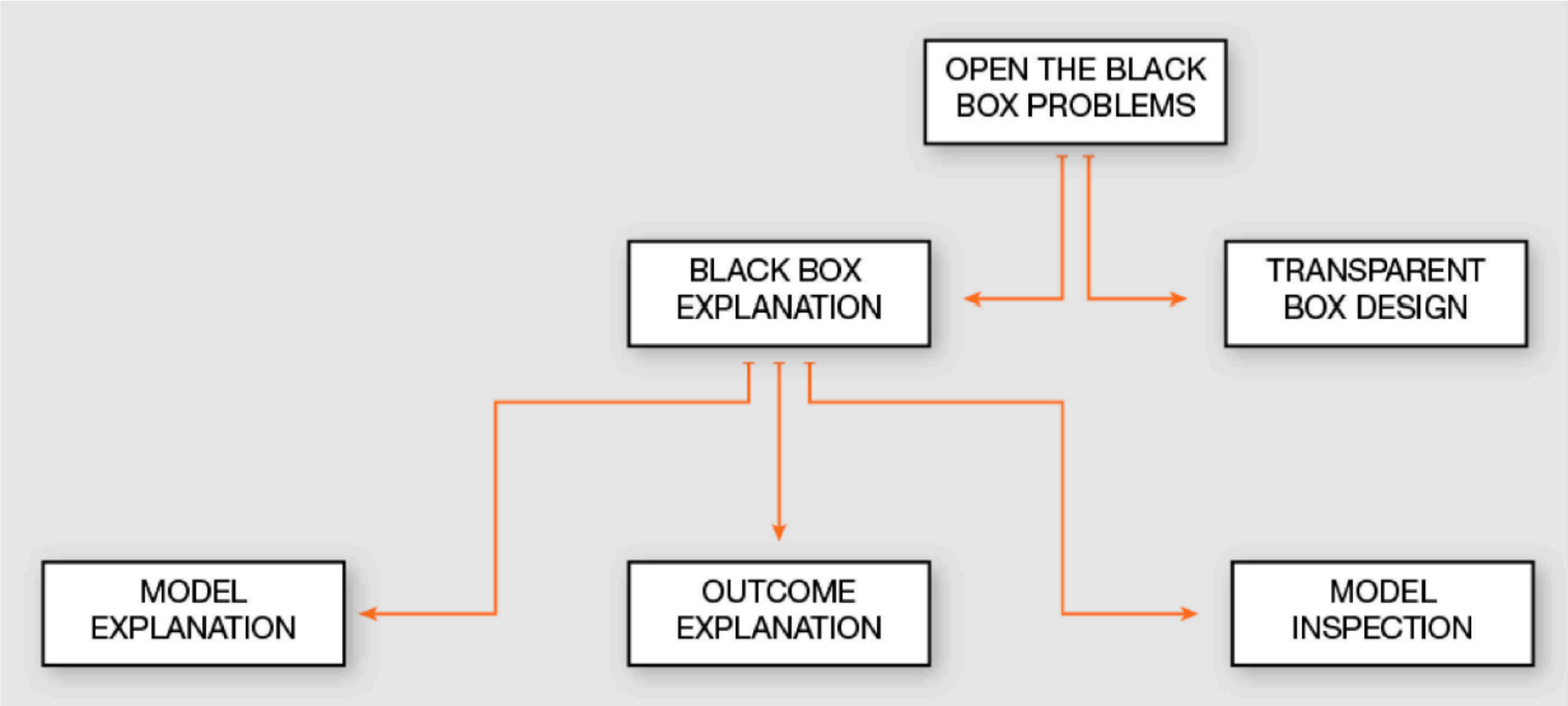
- Linear Model: number of non zero weights in the model.
- Rule: number of attribute-value pairs in condition.
- Decision Tree: estimating the complexity of a tree can be hard.

- 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.

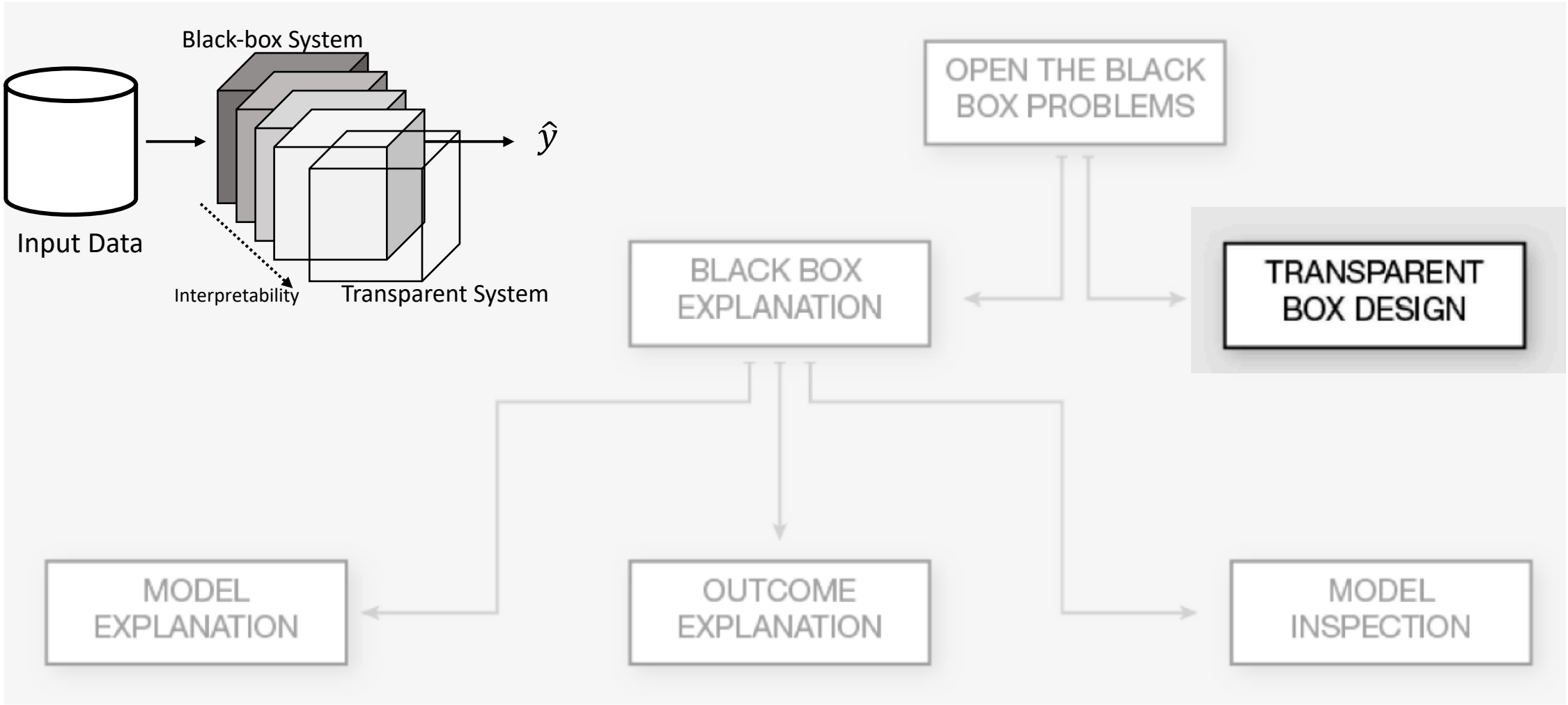


Open the Black Box Problems

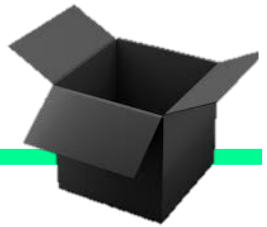
Problems Taxonomy

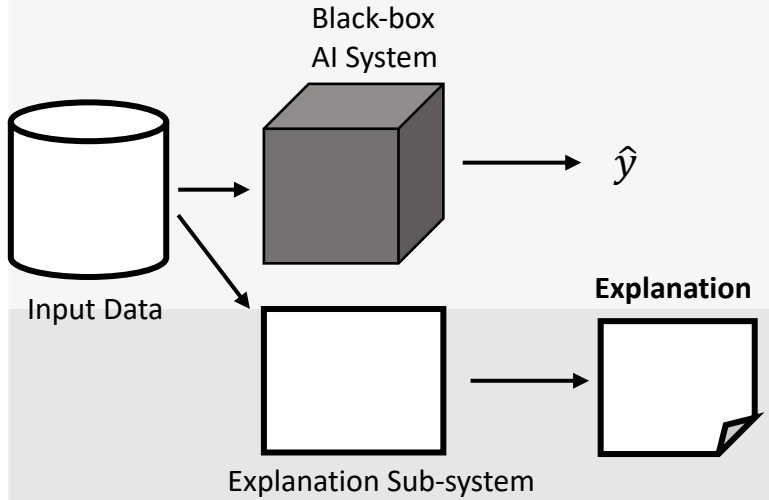


XbD – eXplanation by Design



BBX - Black Box eXplanation





OPEN THE BLACK BOX PROBLEMS

BLACK BOX EXPLANATION

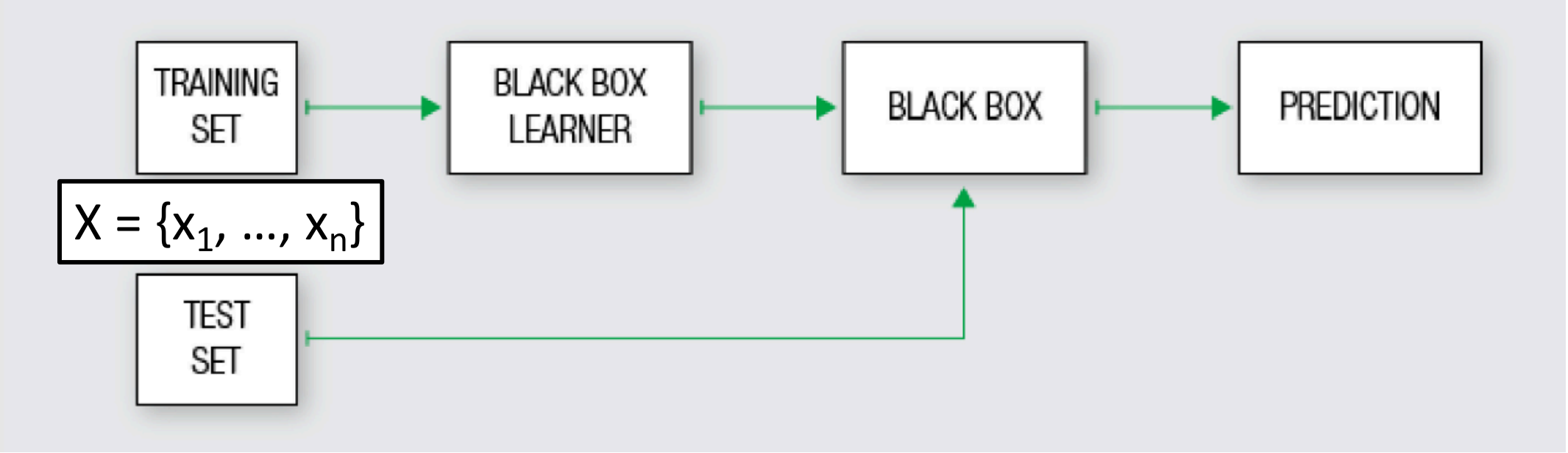
TRANSPARENT BOX DESIGN

MODEL EXPLANATION

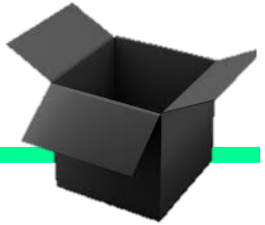
OUTCOME EXPLANATION

MODEL INSPECTION

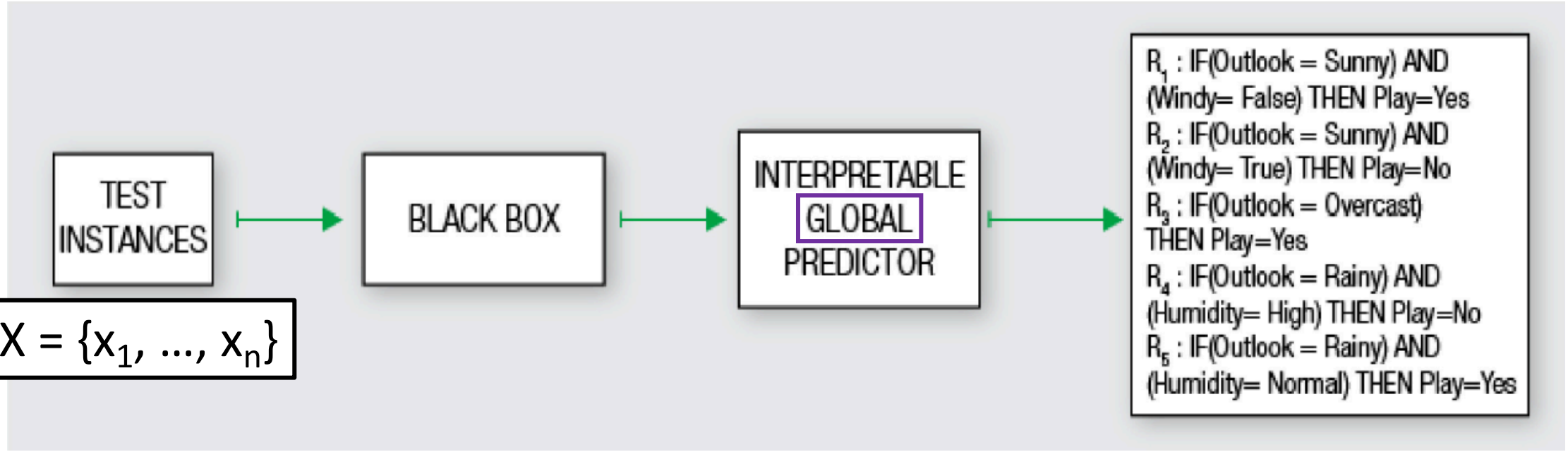
Classification Problem



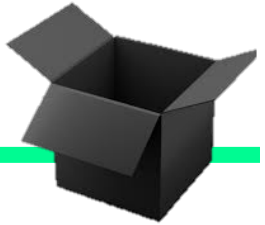
Model Explanation Problem



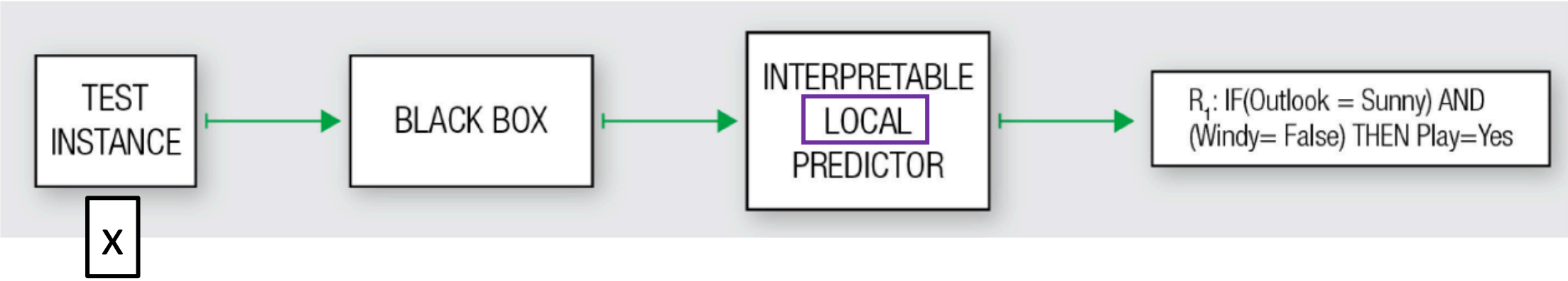
Provide an interpretable model able to mimic the ***overall logic/behavior*** of the black box and to explain its logic**.**



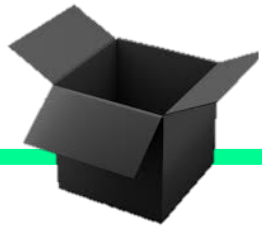
Outcome Explanation Problem



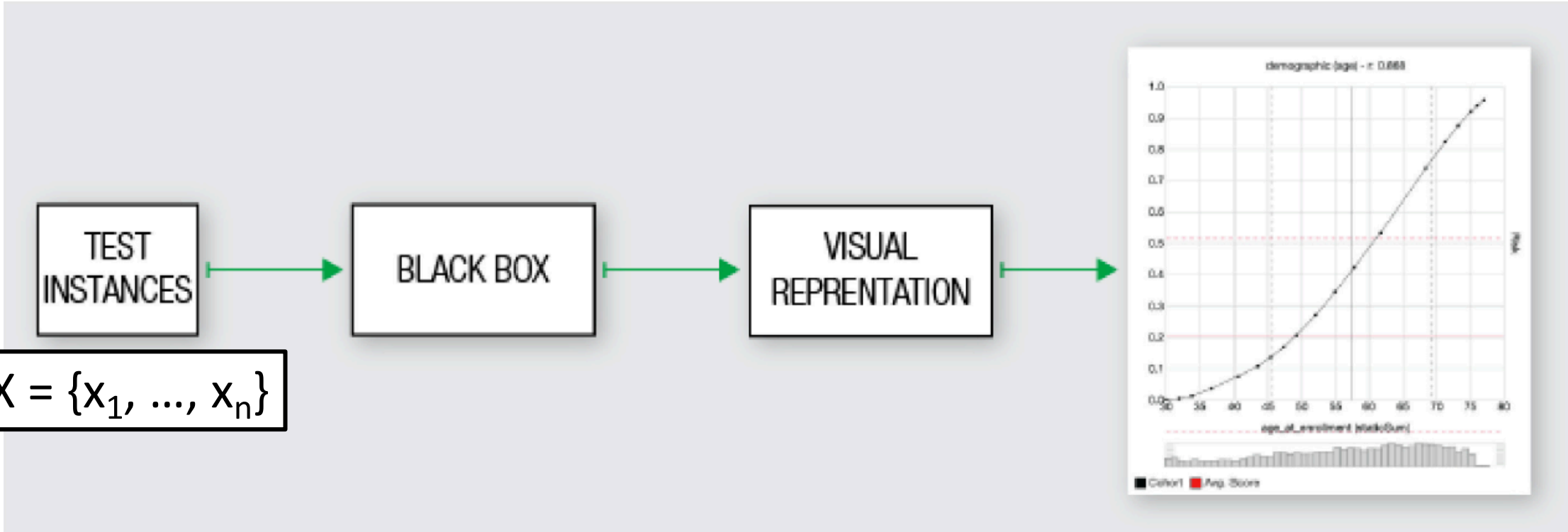
Provide an interpretable outcome, i.e., an ***explanation*** for the outcome of the black box for a ***single instance***.



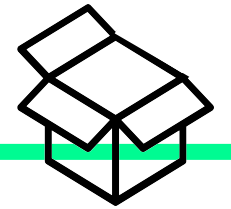
Model Inspection Problem



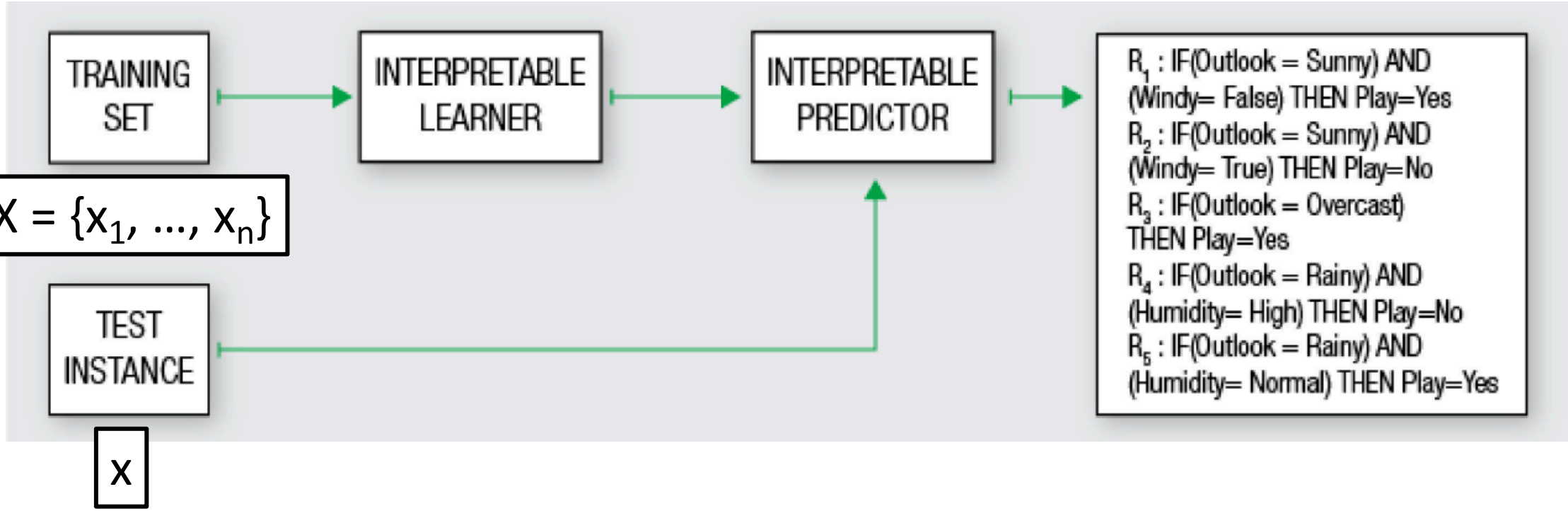
Provide a representation (visual or textual) for understanding either how the black box model works or why the black box returns certain predictions more likely than others.



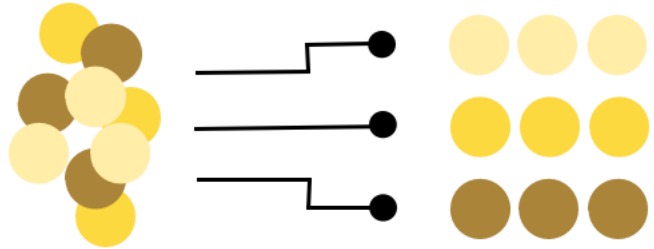
Transparent Box Design Problem



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

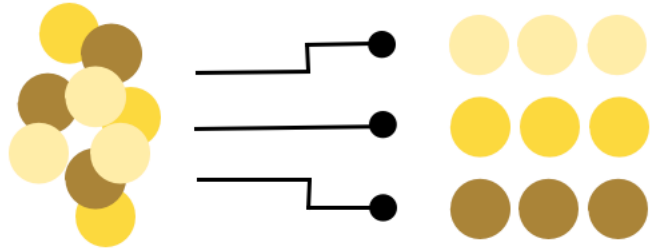


Categorization



- 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

Field names

One row (4 fields)

2000 rows all told

Tabular

(**TAB**)

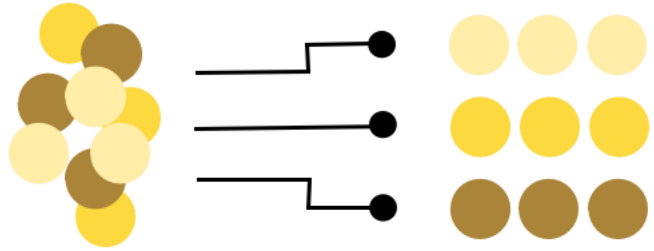
Images

(**IMG**)

Text

(**TXT**)

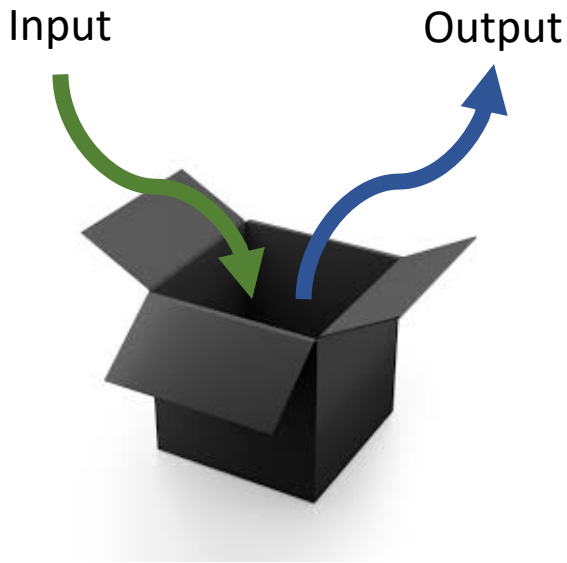
Explanators



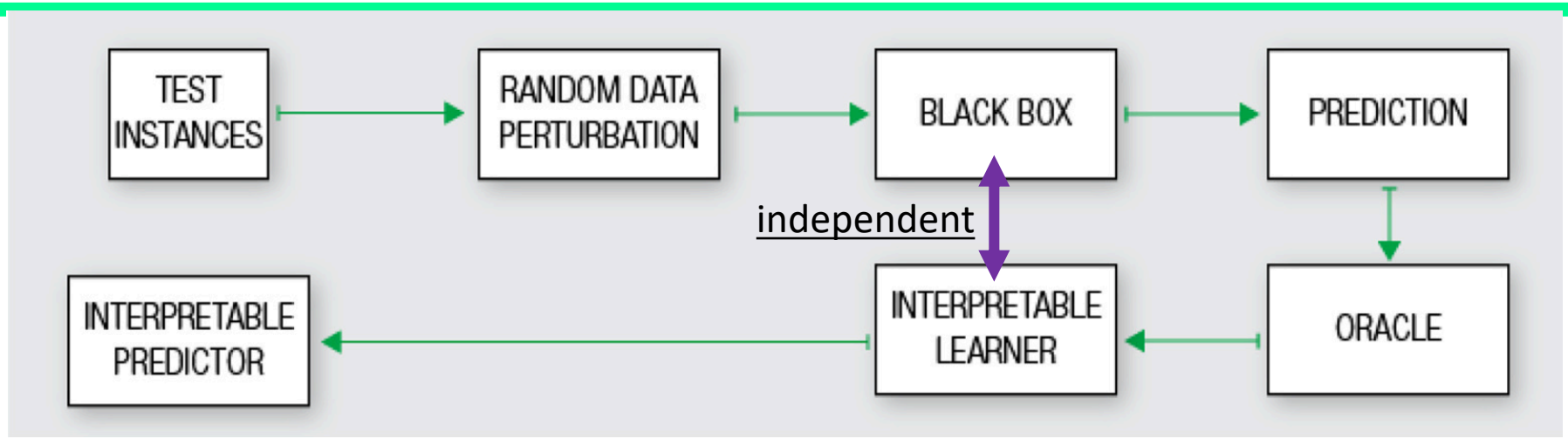
- Decision Tree (***DT***)
- Decision Rules (***DR***)
- Features Importance (***FI***)
- Saliency Maps (***SM***)
- Sensitivity Analysis (***SA***)
- Partial Dependence Plot (***PDP***)
- Prototype Selection (***PS***)
- Activation Maximization (***AM***)

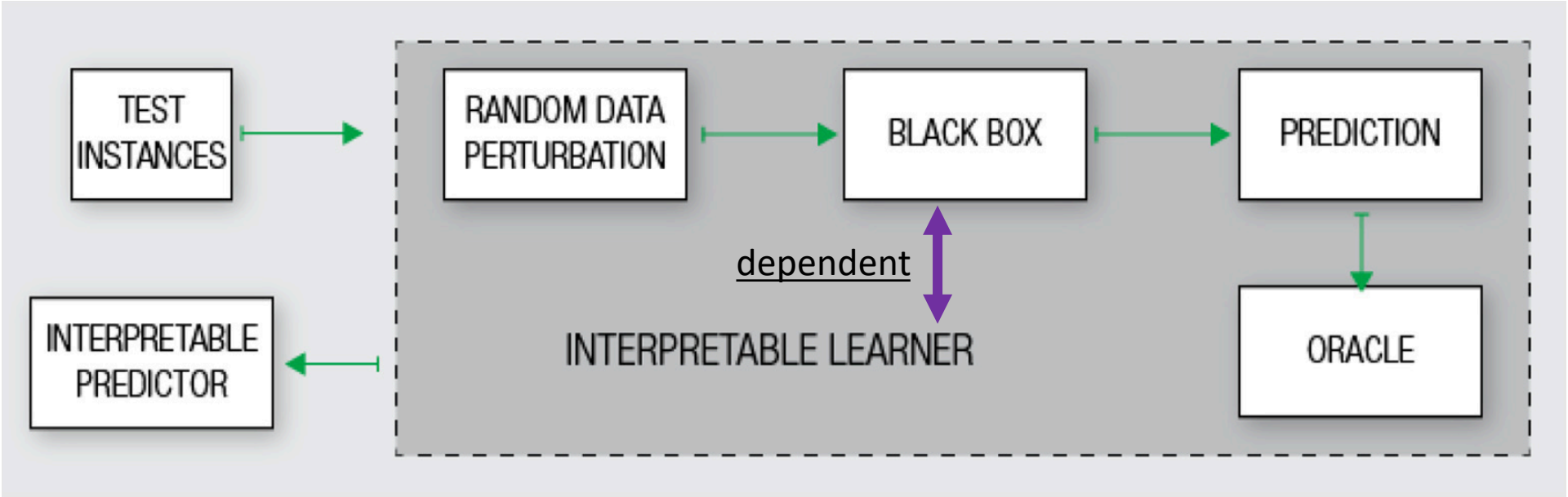
Reverse Engineering

- 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.	1995	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

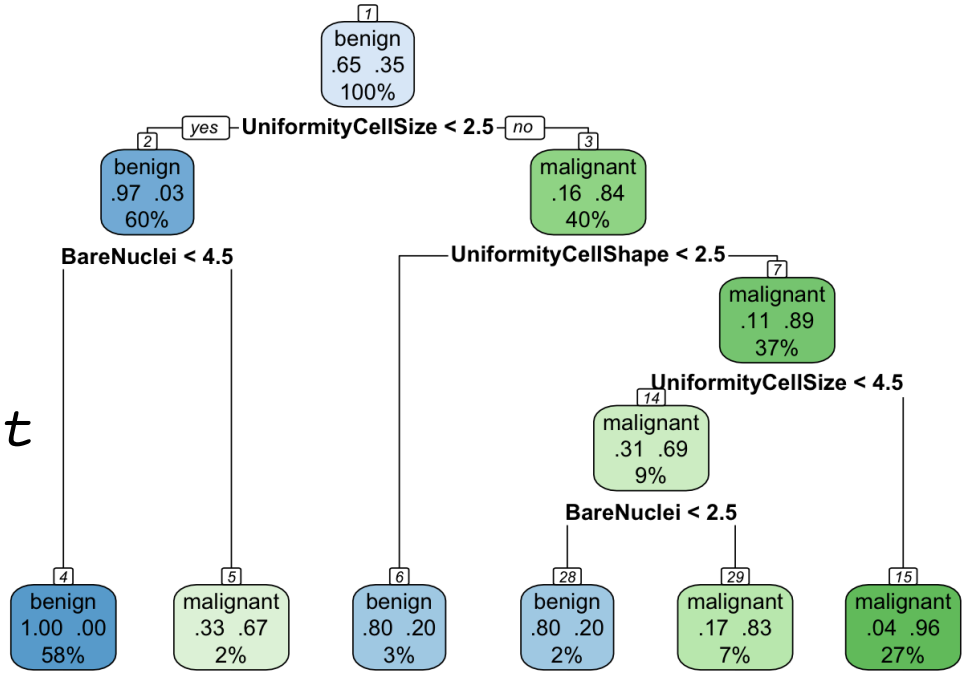
- Explanator: DT
  - Black Box: NN, TE
  - Data Type: TAB
- Explanator: DR
  - Black Box: NN, SVM, TE
  - Data Type: TAB
- Explanator: FI
  - Black Box: AGN
  - Data Type: TAB

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

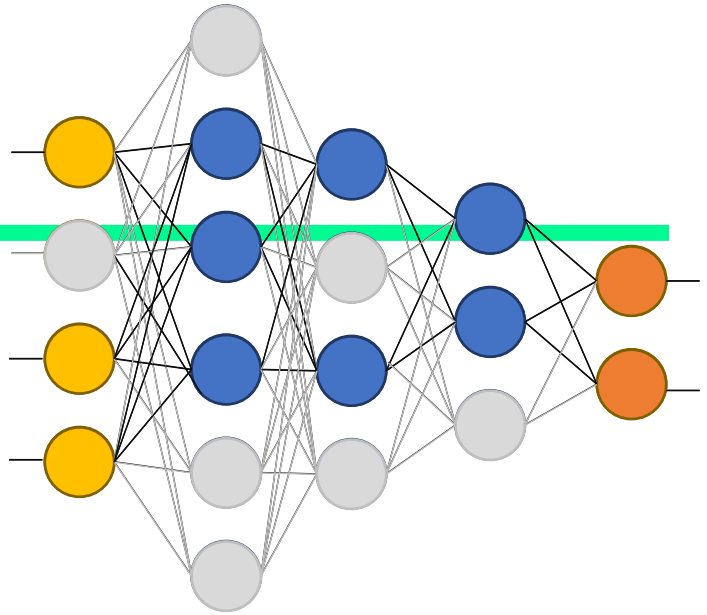
```
01    T = root_of_the_tree()
02    Q = <T, X, {}>
03    while Q not empty & size(T) < limit
04          N, X_N, C_N  = pop(Q)
05          Z_N = random(X_N, C_N)
06          y_Z = b(Z), y = b(X_N)
07          if same_class(y ∪ y_Z)
08                continue
09          S = best_split(X_N ∪ Z_N, y ∪ y_Z)
10          S'= best_m-of-n_split(S)
11          N = update_with_split(N, S')
12          for each condition c in S'
13                C = new_child_of(N)
14                C_C = C_N ∪ {c}
15                X_C = select_with_constraints(X_N, C_N)
16                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



```
01    prune insignificant neurons
02    for each significant neuron
03      for each outcome
04        compute mandatory data ranges
05    for each outcome
06      build rules using data ranges of each neuron
07    prune insignificant rules
08    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.

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

Solving The Outcome Explanation Problem

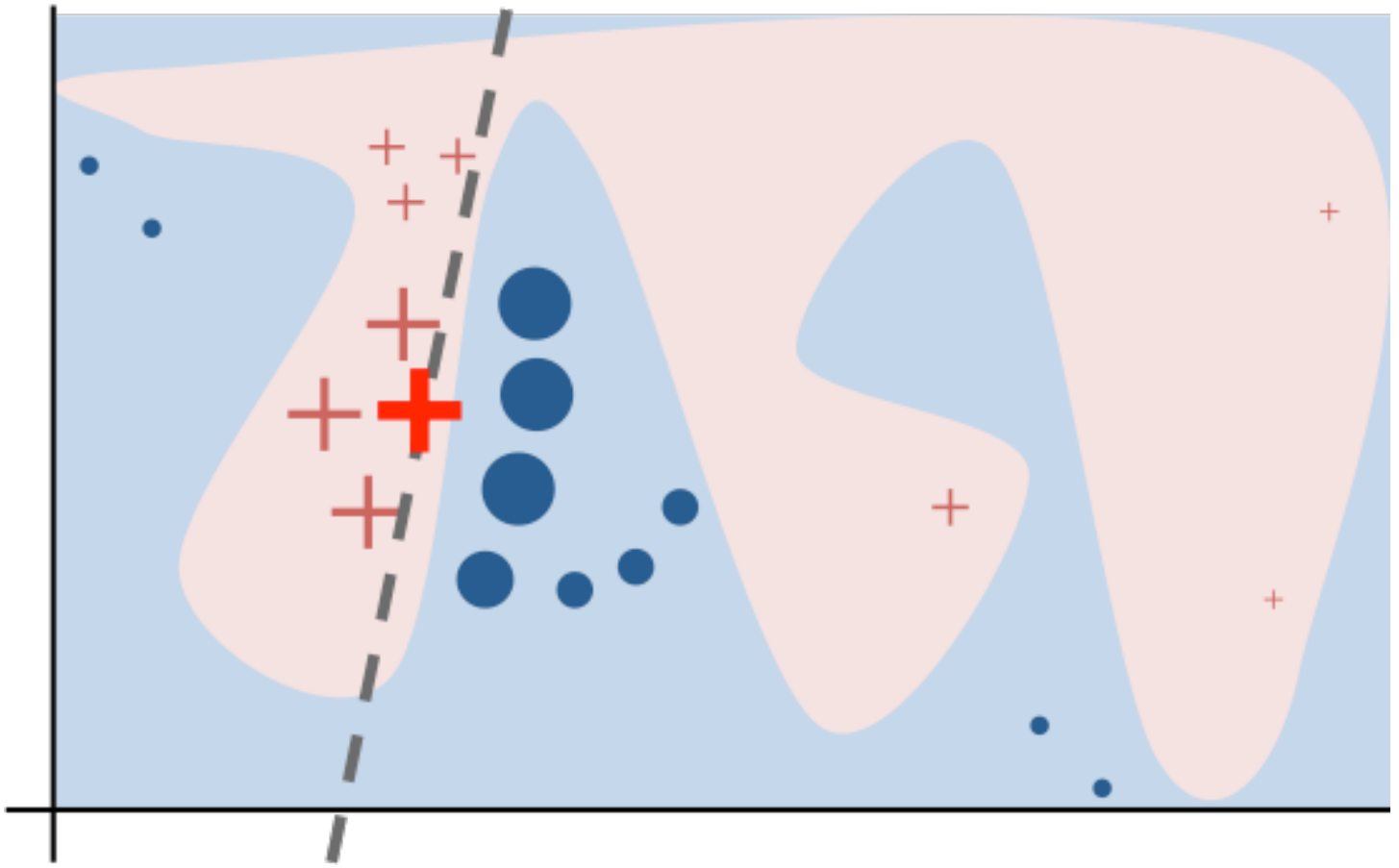
Local Model Explainers

- Explanator: SM
  - Black Box: DNN, NN
  - Data Type: IMG
- Explanator: FI
  - Black Box: DNN, SVM
  - Data Type: ANY
- Explanator: DT
  - Black Box: ANY
  - Data Type: TAB

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

Local Explanation

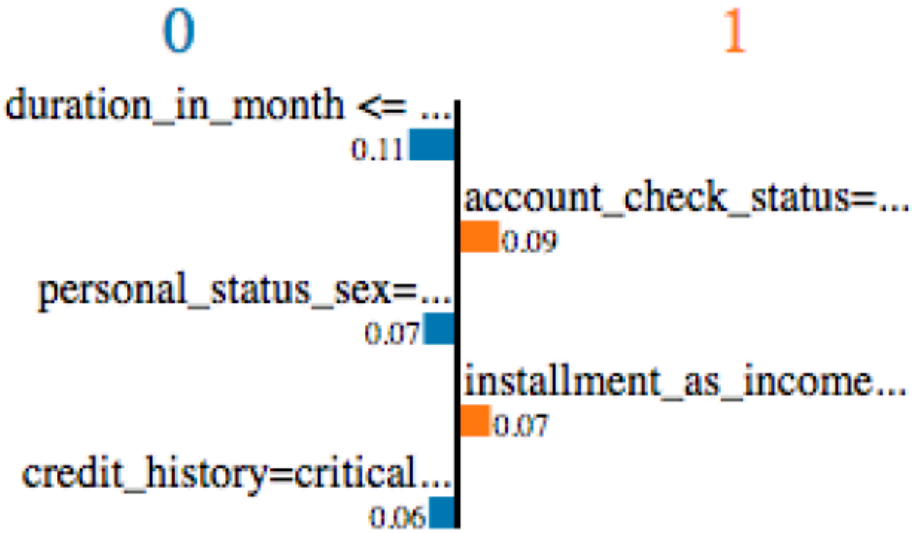
- The overall decision boundary is complex
- In the neighborhood of a single decision, the boundary is simple
- A single decision can be explained by auditing the black box around the given instance and learning a ***local*** decision.

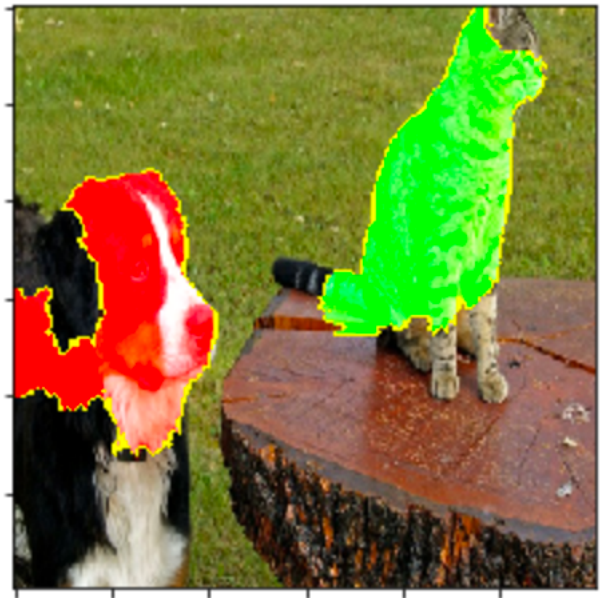


LIME – FI, AGN, ANY

```
01    Z = {}
02    x instance to explain
03    x' = real2interpretable(x)
04    for i in {1, 2, …, N}
05          z_i = sample_around(x')
06          z = interpretabel2real(z')
07          Z = Z U {<z_i, b(z_i), d(x, z)>}
08    w = solve_Lasso(Z, k)
09    return w
```

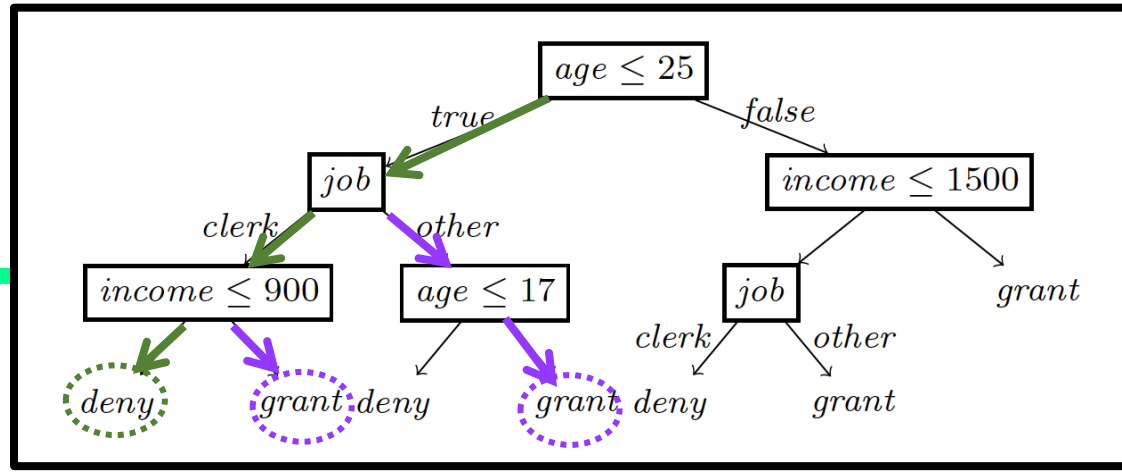
black box auditing





- 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



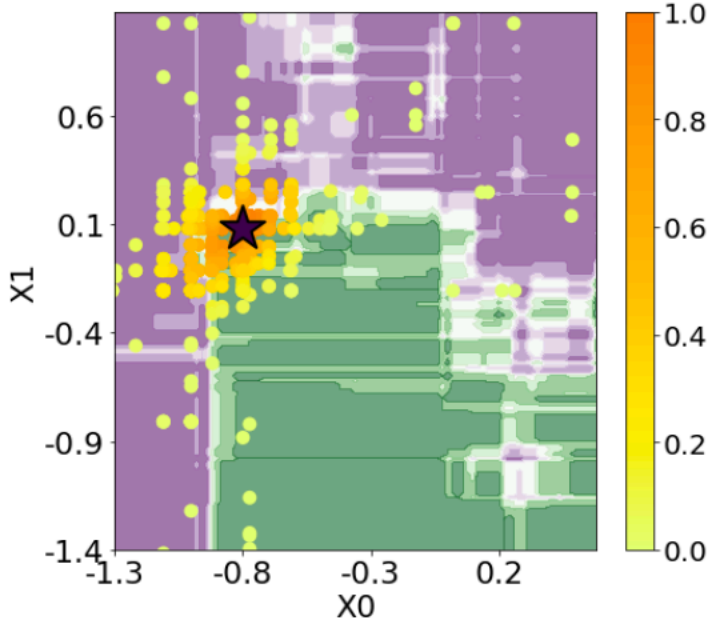

```
01    x instance to explain
02    Z_= = geneticNeighborhood(x, fitness_=, N/2)
03    Z_≠ = geneticNeighborhood(x, fitness_≠, N/2)
04    Z = Z_= ∪ Z_≠
05    c = buildTree(Z, b(Z))
06    r = (p -> y) = extractRule(c, x)
07    ϕ = extractCounterfactual(c, r, x)
08    return e = <r, ϕ>
```

black box auditing

r = {age ≤ 25, job = clerk, income ≤ 900} -> deny

Φ = {({income > 900} -> grant),
({17 ≤ age < 25, job = other} -> grant)}

Pedreschi, Franco Turini,
f black box decision



Meaningful Perturbations – SM, DNN, IMG

01 x instance to explain

02 ***varying*** x into x′ maximizing b(x)~b(x′)

black box auditing

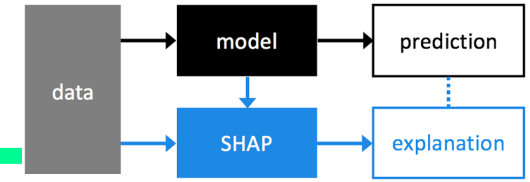
03 the variation runs replacing a region R of x with:

constant value, noise, blurred image

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

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

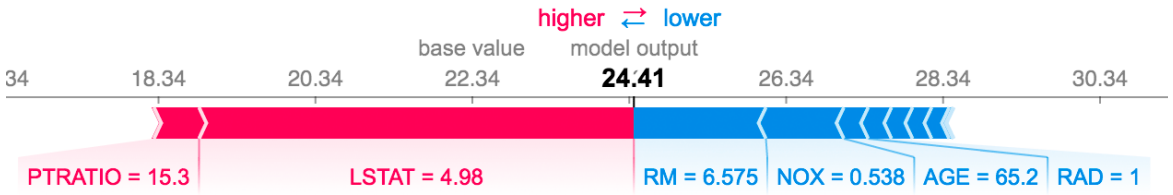
SHAP (SHapley Additive exPlanations)

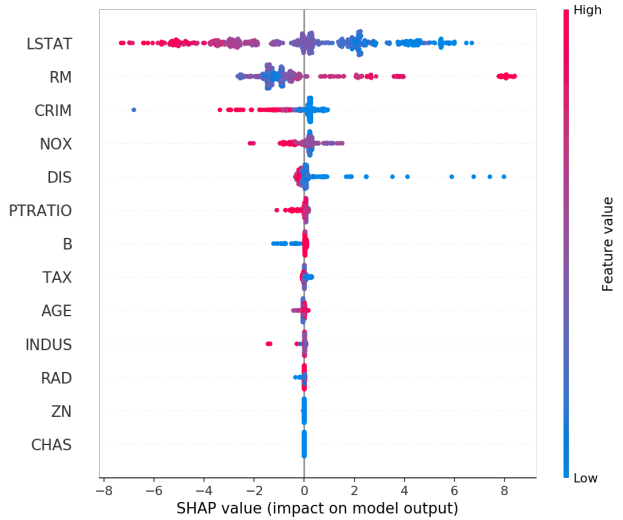


- SHAP assigns each feature an importance value for a particular prediction by means of an additive feature attribution method.
- It assigns an importance value to each feature that represents the effect on the model prediction of including that feature

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

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





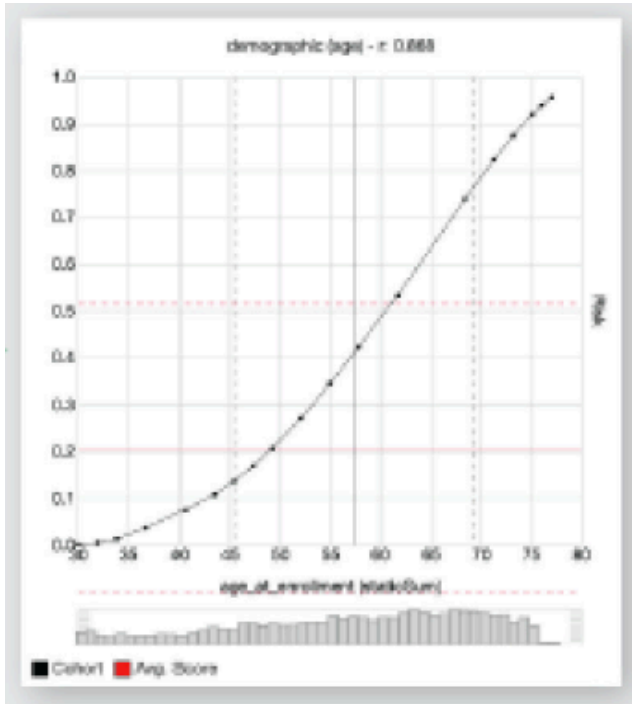
- Lundberg, Scott M., and Su-In Lee. "A unified approach to interpreting model predictions." *Advances in Neural Information Processing Systems*. 2017.

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			✓	✓	✓

Solving The Model Inspection Problem

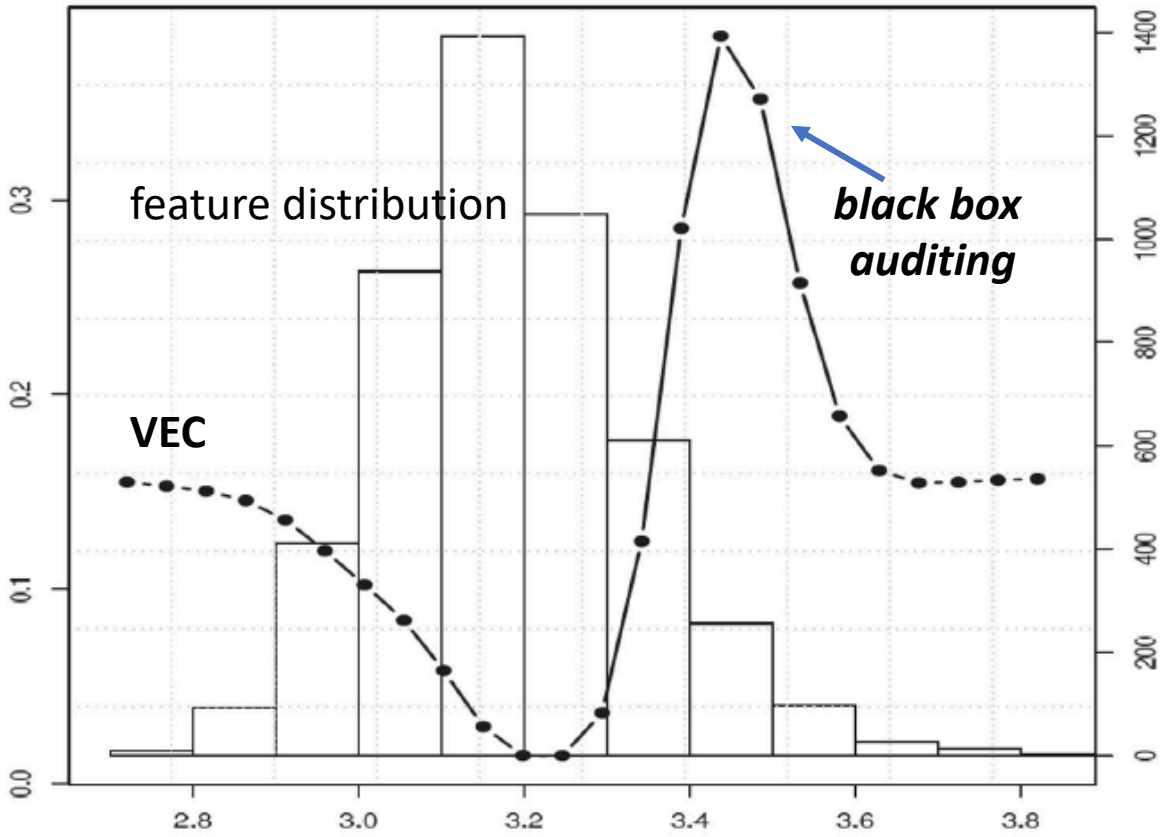
Inspection Model Explainers

- 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

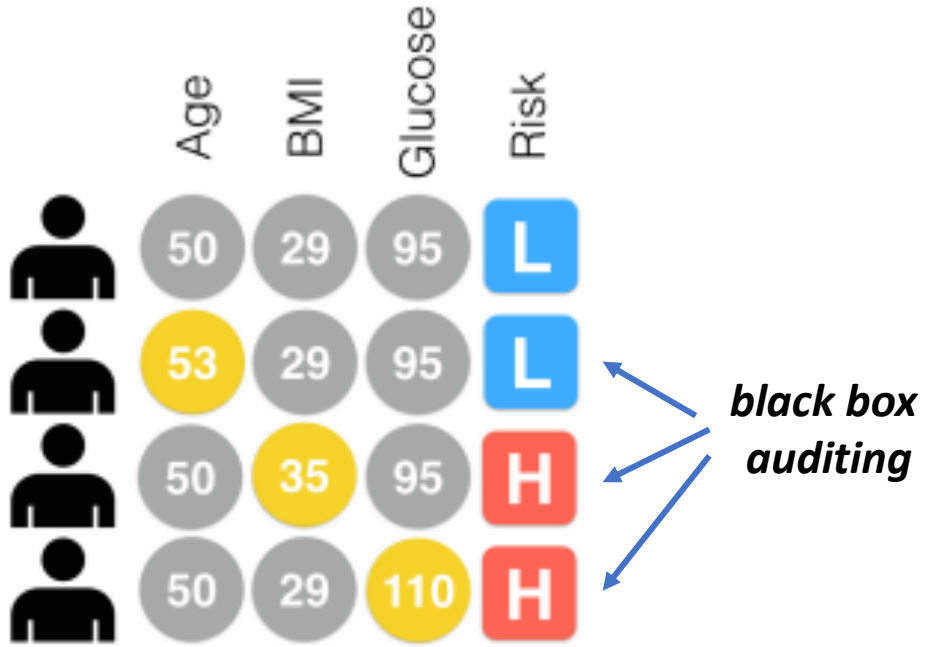
- Sensitivity measures are variables calculated as the range, gradient, variance of the prediction.
- The visualizations realized are barplots for the features importance, and ***Variable Effect Characteristic*** curve (VEC) plotting the input values versus the (average) outcome responses.

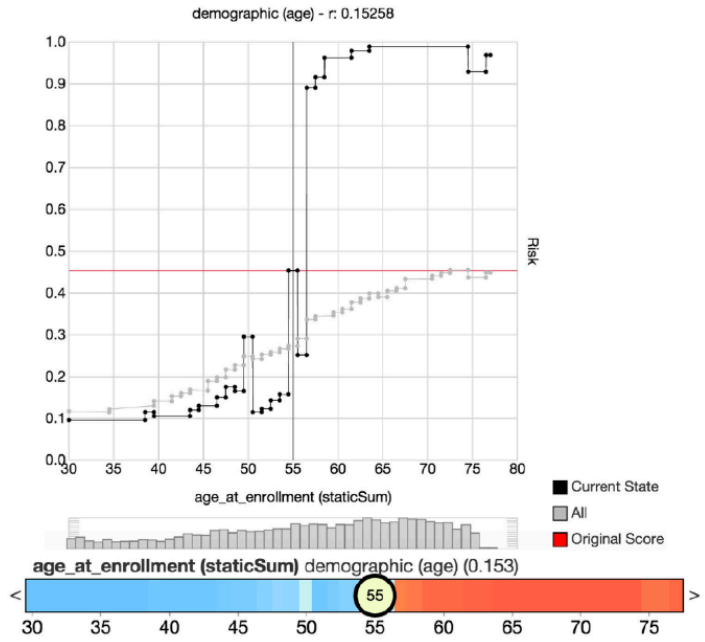


- Paulo Cortez and Mark J. Embrechts. 2011. ***Opening black box data mining models using sensitivity analysis***. CIDM.

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***.





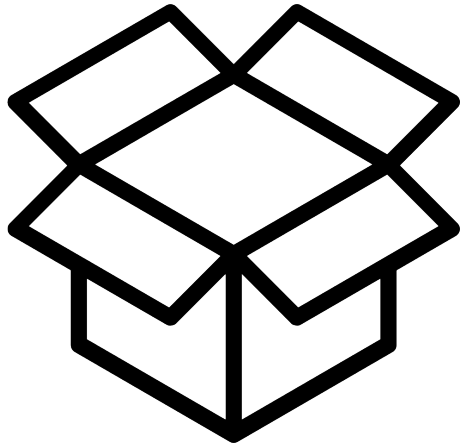
- Ruth Fong and Andrea Vedaldi. 2017. ***Interpretable explanations of black boxes by meaningful perturbation***. arXiv:1704.03296 (2017).

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

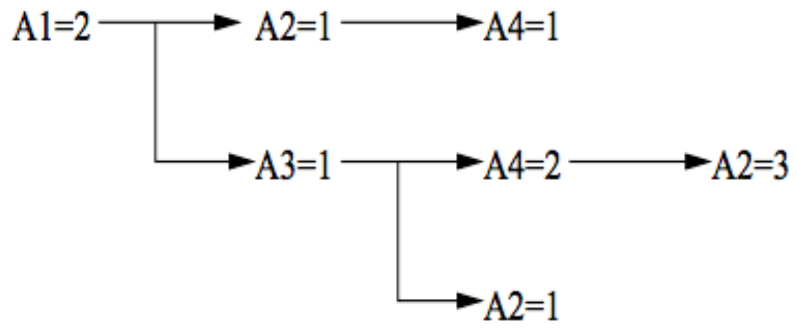
- Explanators:
  - DR
  - DT
  - PS
- Data Type:
  - TAB



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.

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

- 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.

Explaining with Knowledge Graphs

Pasquale Minervini, University College London, London

Outline

- **Knowledge Graphs**
  - What/Where are they?
  - How can they help in XAI?

Outline

- **Knowledge Graphs**
  - What/Where are they?
  - How can they help in XAI?
- **Statistical Relational Learning in Knowledge Graphs**
  - Symbolic (Explainable) Models
  - Sub-Symbolic (Black-Box) Models
  - Incorporating Symbolic Knowledge in Sub-Symbolic Models

Outline

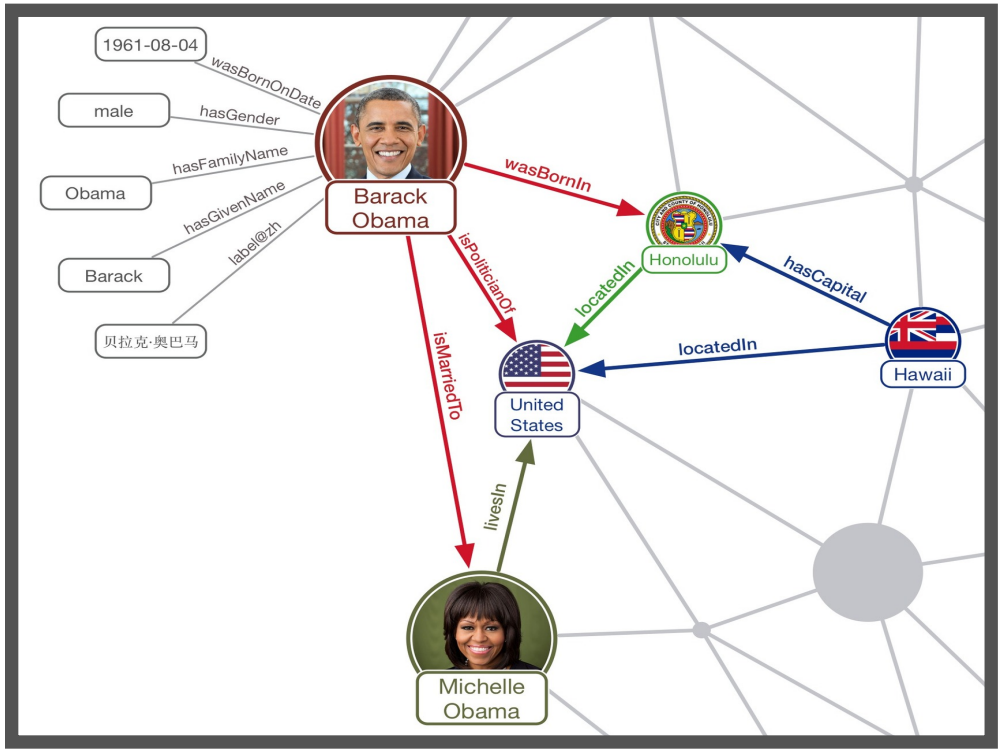
- **Knowledge Graphs**
  - What/Where are they?
  - How can they help in XAI?
- **Statistical Relational Learning in Knowledge Graphs**
  - Symbolic (Explainable) Models
  - Sub-Symbolic (Black-Box) Models
  - Incorporating Symbolic Knowledge in Sub-Symbolic Models
- **Differentiable Reasoning**
  - Forward Chaining
  - Backward Chaining

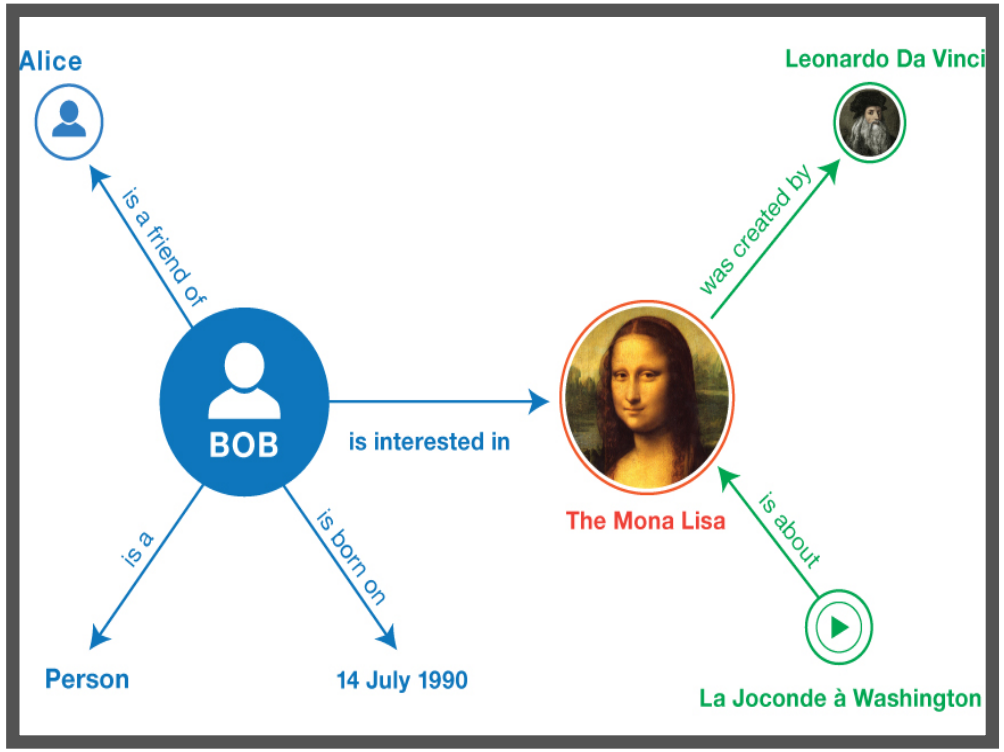
Knowledge Graphs

Knowledge Graphs are *graph-structured Knowledge Bases*, where knowledge is encoded by *relationships between entities.*

Knowledge Graphs

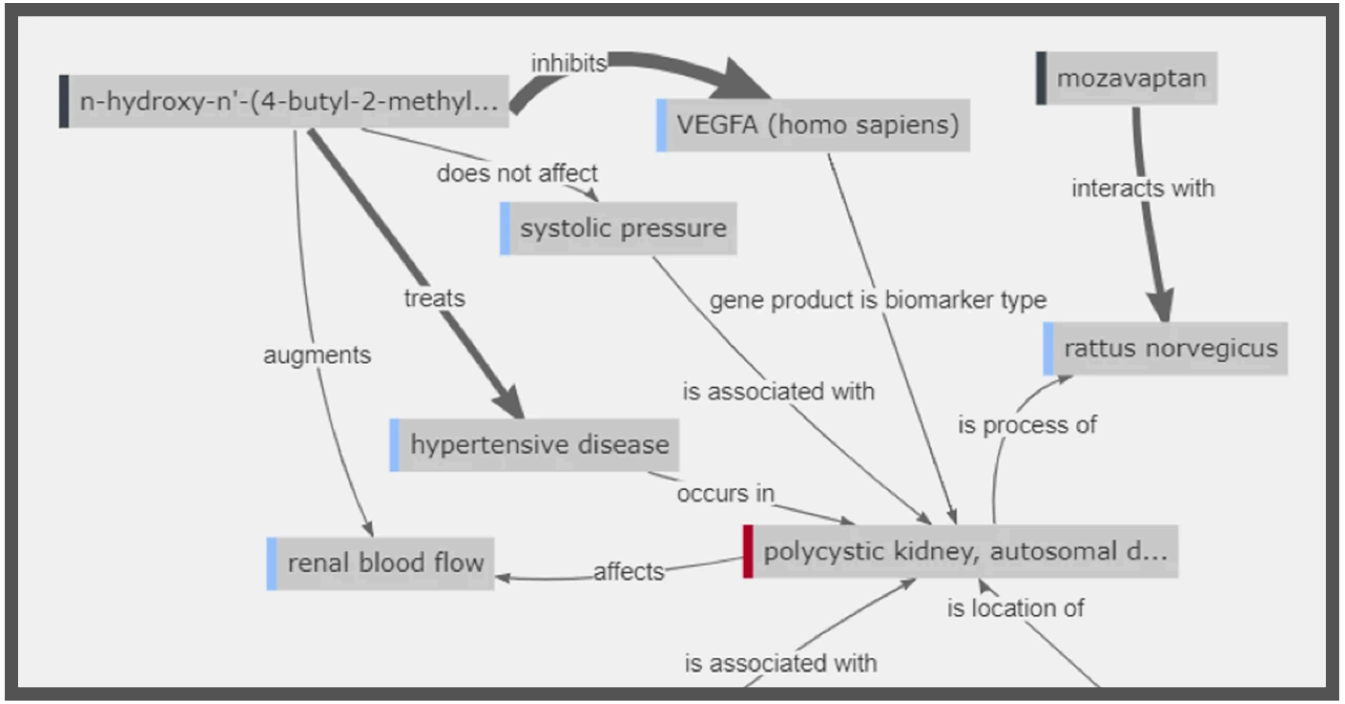
Knowledge Graphs are *graph-structured Knowledge Bases*, where knowledge is encoded by *relationships between entities.*





Knowledge Graphs

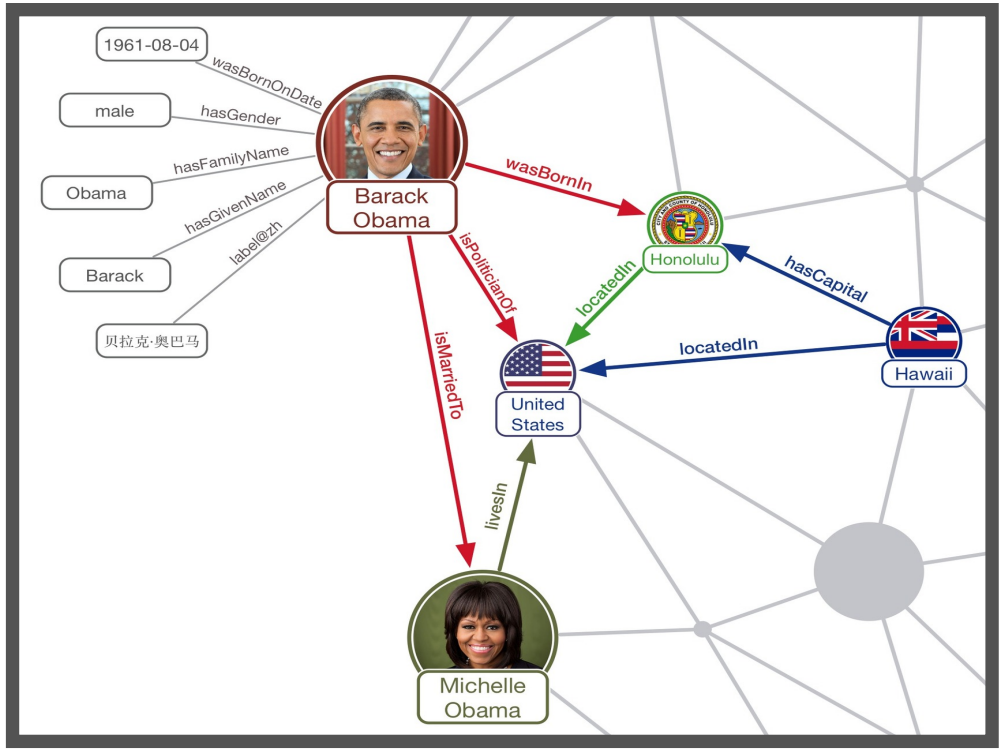
Knowledge Graphs are *graph-structured Knowledge Bases*, where knowledge is encoded by *relationships between entities.*



Drug Prioritization using the semantic properties of a Knowledge Graph, Nature 2019

Knowledge Graphs

Knowledge Graphs are *graph-structured Knowledge Bases*, where knowledge is encoded by *relationships between entities.*

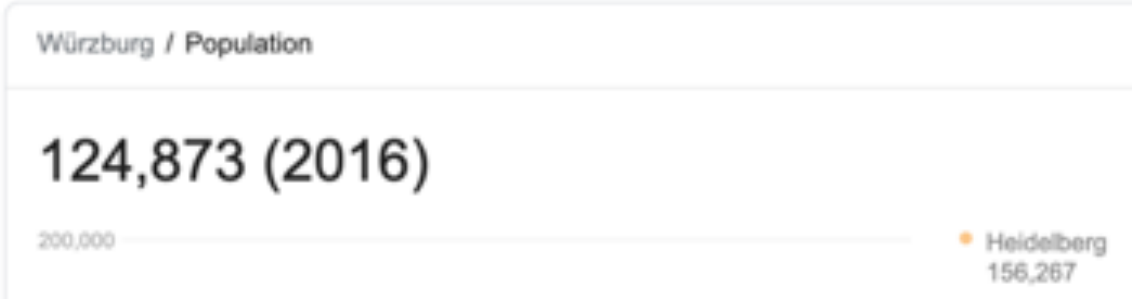


subject	predicate	object
Barack Obama	was born in	Honolulu
Hawaii	has capital	Honolulu
Barack Obama	is politician of	United States
Hawaii	is located in	United States
Barack Obama	is married to	Michelle Obama
Michelle Obama	is a	Lawyer
Michelle Obama	lives in	United States

Industry-Scale Knowledge Graphs in Google

The *Google Knowledge Graph* contains more than 70 billion assertions describing a billion entities and covers a variety of subject matter — “things not strings”.

Used for answering factoid queries about entities served from the Knowledge Graph.



1 Billion entities, ~70 Billion assertions

Würzburg

City in Bavaria

Wurzburg is a city in Germany's Bavaria region. It's famous for lavish baroque and rococo architecture, particularly the 18th-century palace residence, with ornate rooms, a huge fresco by Venetian artist Tiepolo and an elaborate staircase. Home to numerous wine bars, cellars and wineries, Wurzburg is the center of the Franconian wine country, with its distinctive bocksbeutel (bottles with flattened round shapes).

Weather : 4 ° C, Wind N at 3 km / h, 82% Humidity

Postal codes : 97070-97084

District : Urban district

Population : 124,873 (2016) United Nations

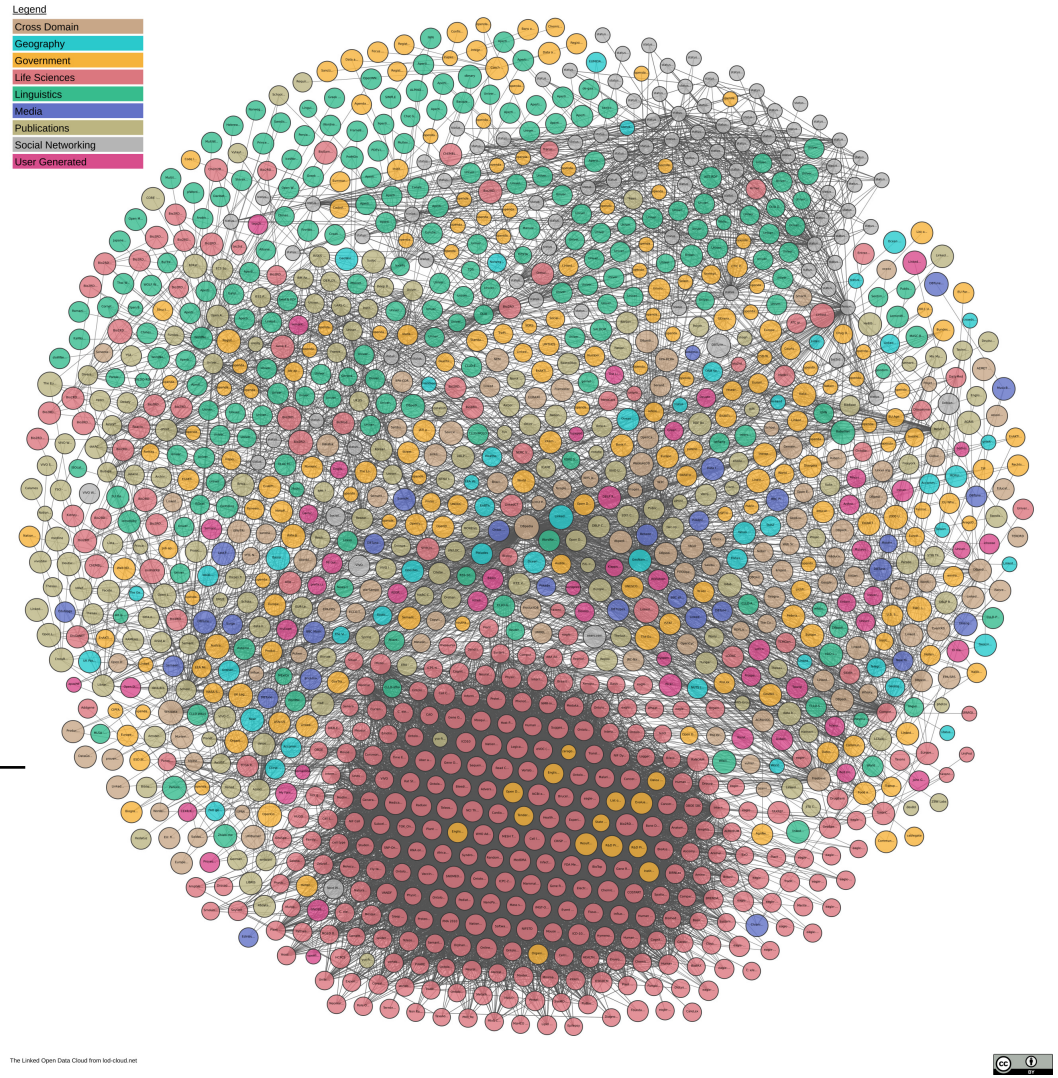
Dialing code : 0931

The Linked Open Data Cloud

Linked Open Data cloud - over 1200 interlinked KGs encoding more than 200M facts about more than 50M entities.

Spans a variety of domains, such as Geography, Government, Life Sciences, Linguistics, Media, Publications, and Cross-domain

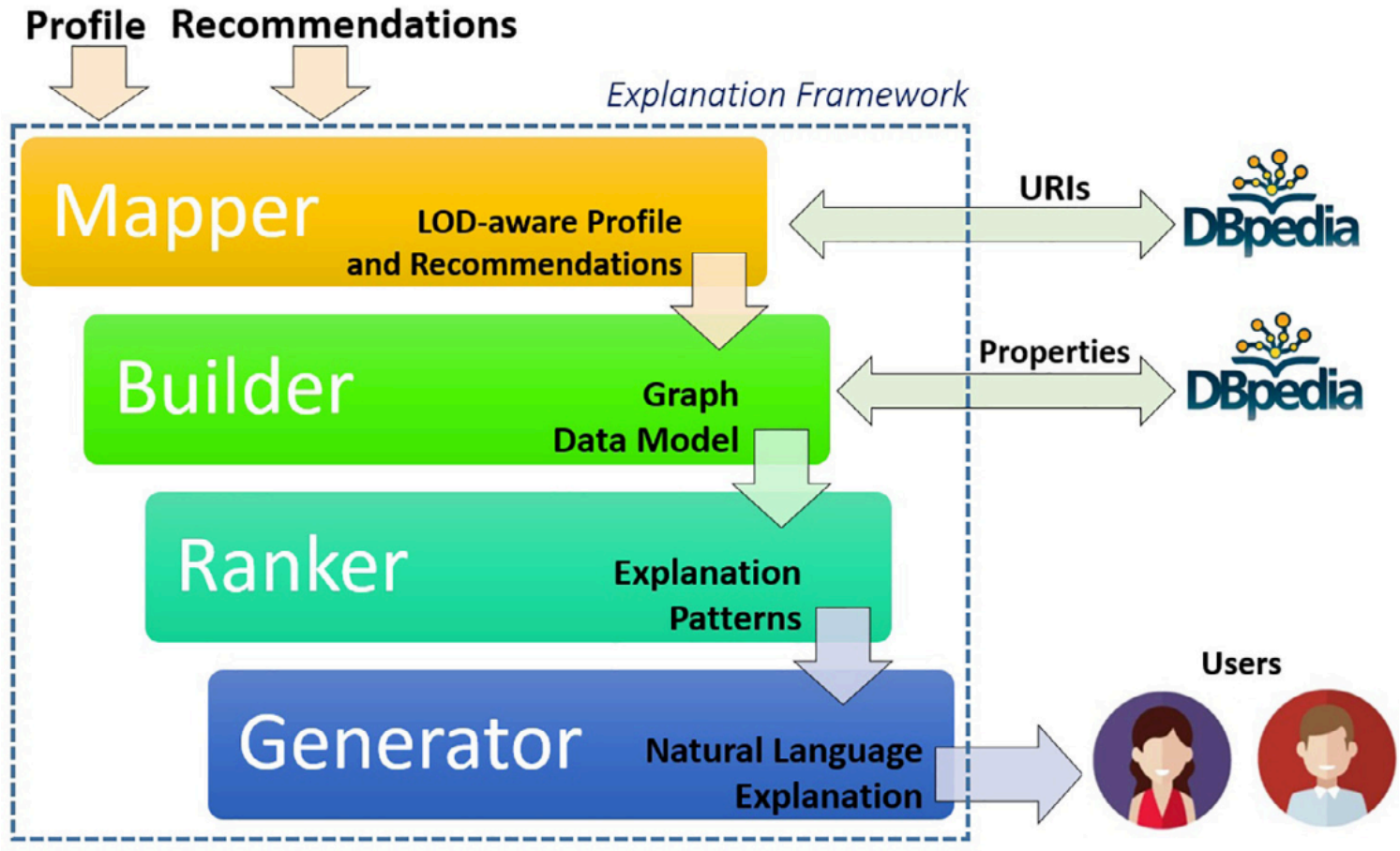
Name	Entities	Relations	Types	Facts
Freebase	40M	35K	26.5K	637M
DBpedia (en)	4.6M	1.4K	735	580M
YAGO3	17M	77	488K	150M
Wikidata	15.6M	1.7K	23.2K	66M



Knowledge Graphs and Explainable AI

We can use Knowledge Graphs for *explaining* the decisions of Machine Learning algorithms, such as recommender systems, and design machine learning models that are less prone to capturing *spurious correlations* in the data.

- Locally vs. Globally
- Ad-hoc vs. Post-hoc

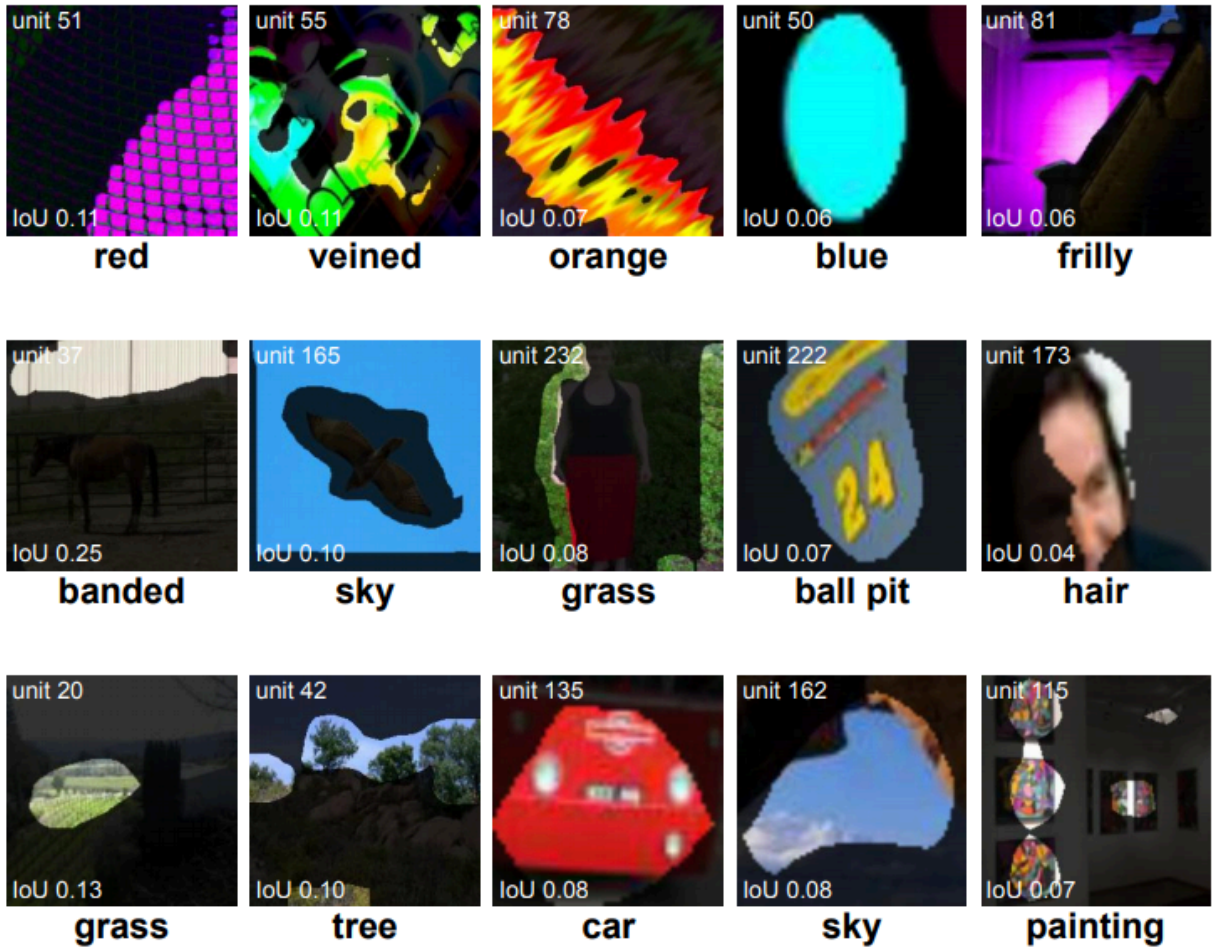


[Di Noia et al. 2012, Ostuni et al. 2013, Musto et al. 2019]

Knowledge Graphs and Explainable AI

We can use Knowledge Graphs for *explaining* the decisions of Machine Learning algorithms, such as recommender systems, and design machine learning models that are less prone to capturing *spurious correlations* in the data.

- Locally vs. Globally
- Ad-hoc vs. Post-hoc



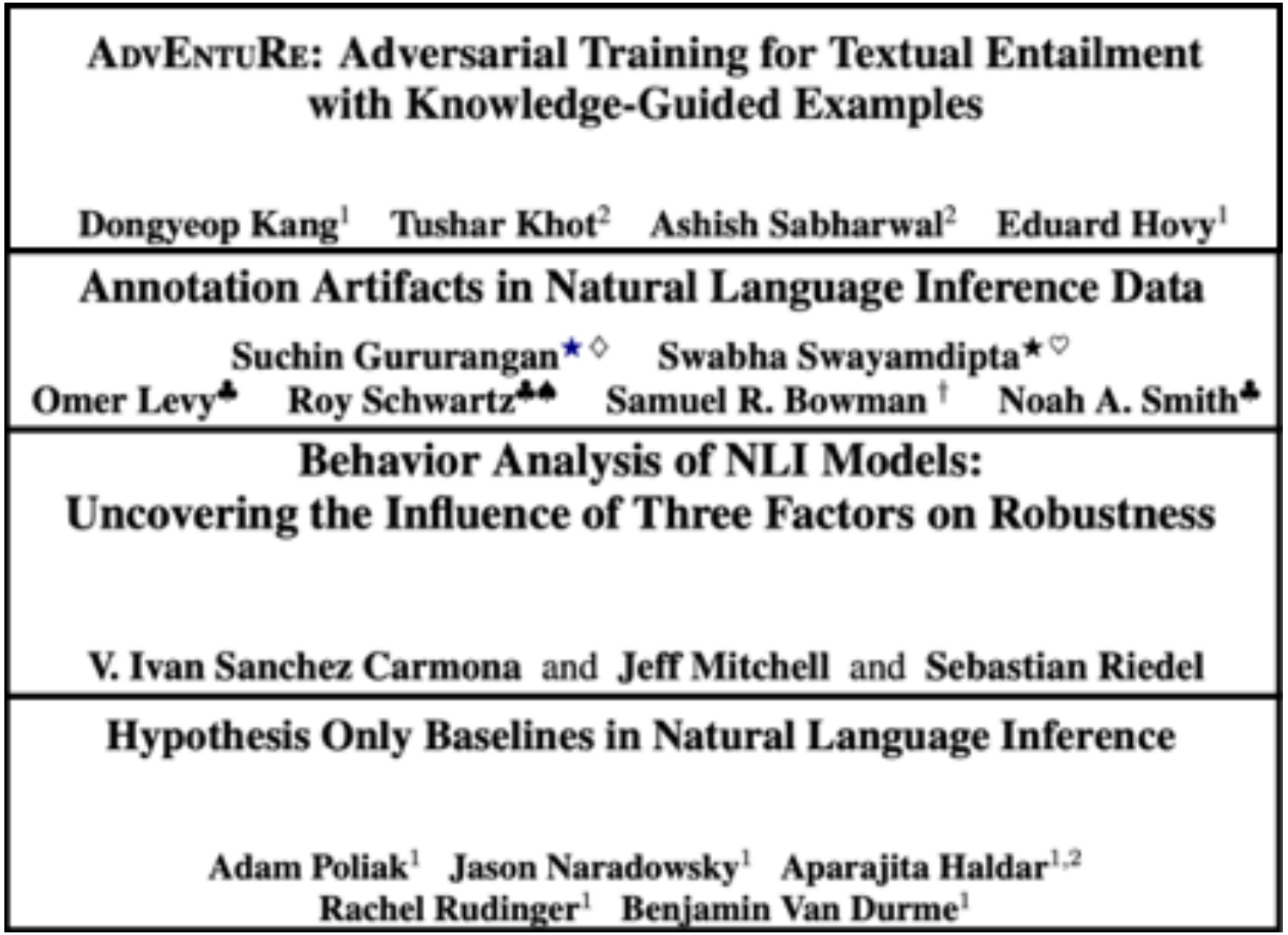
Freddy Lecue: On The Role of Knowledge Graphs in Explainable AI. SWJ 2019

[Bau et al. 2017, Lecue 2019]

Knowledge Graphs and Explainable AI

We can use Knowledge Graphs for *explaining* the decisions of Machine Learning algorithms, such as recommender systems, and design machine learning models that are less prone to capturing *spurious correlations* in the data.

- Locally vs. Globally
- Ad-hoc vs. Post-hoc



AdvEntuRe: Adversarial Training for Textual Entailment with Knowledge-Guided Examples

Dongyeop Kang[1] Tushar Khot[2] Ashish Sabharwal[2] Eduard Hovy[1]

Annotation Artifacts in Natural Language Inference Data

Suchin Gururangan[★◇] Swabha Swayamdipta[★♡]
Omer Levy[♣] Roy Schwartz[♣♠] Samuel R. Bowman[†] Noah A. Smith[♣]

Behavior Analysis of NLI Models: Uncovering the Influence of Three Factors on Robustness

V. Ivan Sanchez Carmona and Jeff Mitchell and Sebastian Riedel

Hypothesis Only Baselines in Natural Language Inference

Adam Poliak[1] Jason Naradowsky[1] Aparajita Haldar[1,2]
Rachel Rudinger[1] Benjamin Van Durme[1]

Statistical Relational Learning

- **Task** — model the existence of each triple $x_{spo} = (s, p, o) \in \mathscr{E} \times \mathscr{R} \times \mathscr{E}$ as *binary random variables* $y_{spo} \in \{0,1\}$ indicating whether x_{spo} is in the KG:

$$y_{spo} = \begin{cases} 1 & \text{if } x_{spo} \in \mathscr{G} \\ 0 & \text{otherwise} \end{cases} \quad \text{entries in} \quad \overline{\mathbf{Y}} \in \{0,1\}^{|\mathscr{E}| \times |\mathscr{R}| \times |\mathscr{E}|}$$

- Every realisation of $\overline{\mathbf{Y}}$ denotes a *possible world*-modelling $P\left(\overline{\mathbf{Y}}\right)$ allows predicting triples based on the state of the entire Knowledge Graph.
- Scalability is important - e.g. on Freebase (40M entities), the number of variables to represent can be quite large: $|\mathscr{E} \times \mathscr{R} \times \mathscr{E}| > 10^{19}$

Types of Statistical Relational Learning Models

- Depending on our assumptions on $P(\overline{\mathbf{Y}})$, we end up with *three model classes*:
- **Latent Feature Models**: variables $y_{spo} \in \{0,1\}$ are *conditionally independent* given the *latent features* $\boldsymbol{\Theta}$ associated with subject, predicate, and object:

$$\forall x_i, x_j \in \mathscr{E} \times \mathscr{R} \times \mathscr{E}, x_i \neq x_j : y_i \perp\!\!\!\perp y_j \mid \boldsymbol{\Theta}$$

- **Observable Feature Models**: related to Latent Feature Models, but $\boldsymbol{\Theta}$ are now *graph-based features*, such as *paths* linking the subject and the object.
- **Graphical Models**: variables $y_{spo} \in \{0,1\}$ are not assumed to be conditionally independent — each y_{spo} can depend on any of the other random variables in $\overline{\mathbf{Y}}$.

Conditional Independence Assumption

- Assuming all y_{spo} variables are conditionally independent allows modelling their existence via a scoring function $f(s,p,o \mid \Theta)$ representing the likelihood that a triple is in the KG, conditioned on the parameters $\boldsymbol{\Theta}$:

$$P(\overline{\mathbf{Y}} \mid \Theta) = \prod_{s \in \mathscr{E}} \prod_{p \in \mathscr{R}} \prod_{o \in \mathscr{E}} \begin{cases} P\left(y_{spo} \mid \Theta\right) & \text{if } y_{spo} = 1 \\ 1 - P\left(y_{spo} \mid \Theta\right) & \text{otherwise} \end{cases} \quad \text{with } P\left(y_{spo} \mid \Theta\right) = \sigma\left(f(s,p,o \mid \Theta)\right)$$

- **Scoring Function** - depending on the type of features used by $f(\cdot \mid \Theta)$ we have two families of models - *Observable* and *Latent Feature Models*.

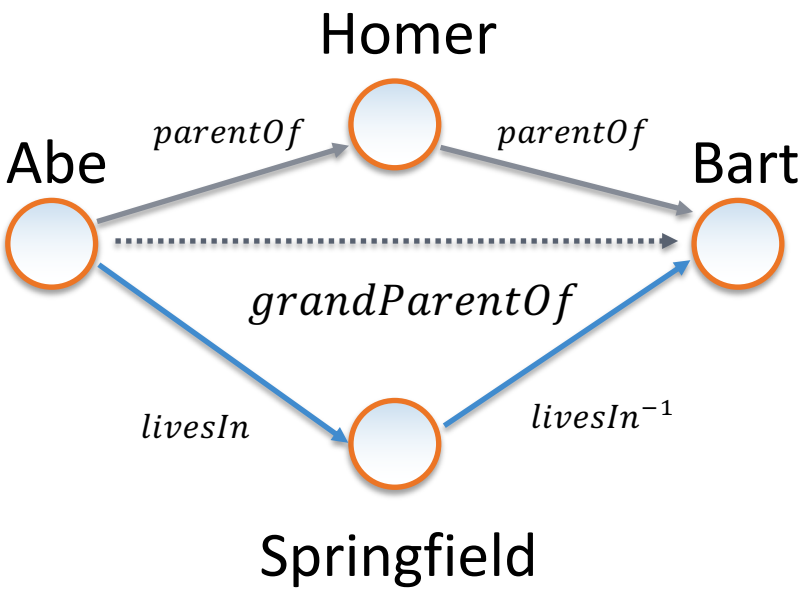
Observable Feature Models - Rule Mining and ILP

Rule Mining and **Inductive Logic Programming** methods extract rules via mining methods, and use them to infer new links.

- **Logic Programming (deductive):** from facts and rules, infer new facts (First-Order Logic)
- **Inductive Logic Programming (ILP):** from correlated facts, infer new rules **(e.g. Progol** [Muggleton, 1993]**, Aleph** [Srinivasan, 1999]**, DL-Learner** [Lehmann, 2009]**, FOIL** [Quinlan, 1990]**, ..)**
- **Rule Mining:** AMIE [Galárraga et al. 2015] is orders of magnitude faster than traditional ILP methods, and consistent with the Open World Assumption in Knowledge Graphs:
  - Partial Completeness Assumption
  - Efficient search space exploration via Mining Operators

Observable Feature Models - Path Ranking Algorithm

Path Ranking Algorithm (PRA) uses *length-bounded random walks* as features between entity pairs for predicting a target relation [Lao et al. 2010].



A **PRA model** scores a subject-object pair by a linear function of their path features:

$$f(s,p,o) = \sum_{\pi \in \Pi_p} P(s \to o \mid \pi) \times \theta_{\pi,p}$$

where Π is the set of all length-bounded relation paths, and θ are parameters estimated via L1,L2-regularised logistic regression.

Some extensions: Subgraph Features [Gardner et al. 2015], Multi-Task [Wang et al. 2016]

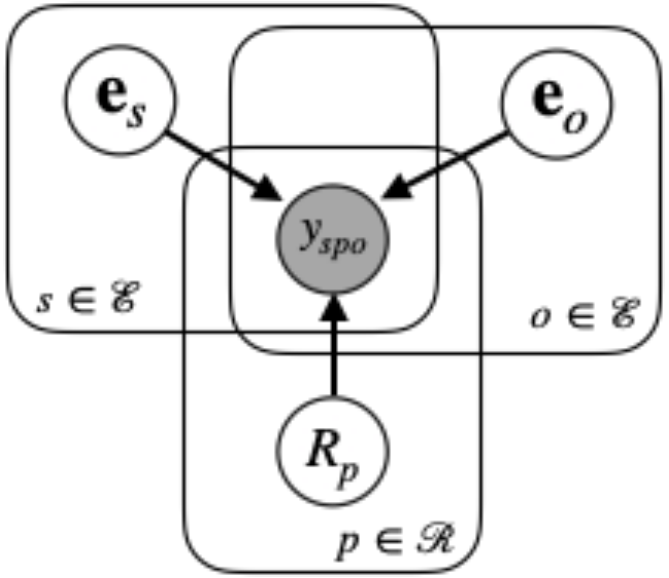
Observable Feature Models are Interpretable

Rules extracted by AMIE+ [Galárraga et al. 2015] from the YAGO3-10 dataset [Dettmers et al. 2018]

Body	$\Rightarrow$ *Head*	*Confidence*
$hasNeighbor(X,Y)$	$\Rightarrow hasNeighbor(Y,X)$	0.99
$isMarriedTo(X,Y)$	$\Rightarrow isMarriedTo(Y,X)$	0.96
$hasNeighbor(X,Z) \wedge hasNeighbor(Z,Y)$	$\Rightarrow hasNeighbor(X,Y)$	0.88
$isAffiliatedTo(X,Y)$	$\Rightarrow playsFor(Y,X)$	0.87
$playsFor(X,Y)$	$\Rightarrow isAffiliatedTo(Y,X)$	0.75
$dealsWith(X,Z) \wedge dealsWith(Z,Y)$	$\Rightarrow dealsWith(X,Y)$	0.73
$isConnectedTo(X,Y)$	$\Rightarrow isConnectedTo(Y,X)$	0.66
$dealsWith(X,Z) \wedge imports(Z,Y)$	$\Rightarrow imports(X,Y)$	0.61
$influences(Z,X) \wedge isInterestedIn(Z,Y)$	$\Rightarrow isInterestedIn(X,Y)$	0.53

Latent Feature Models

- Variables y_{spo} are conditionally independent given a set of latent features and parameters Θ. Latent means that are not directly observed in the data, and thus need to be estimated.

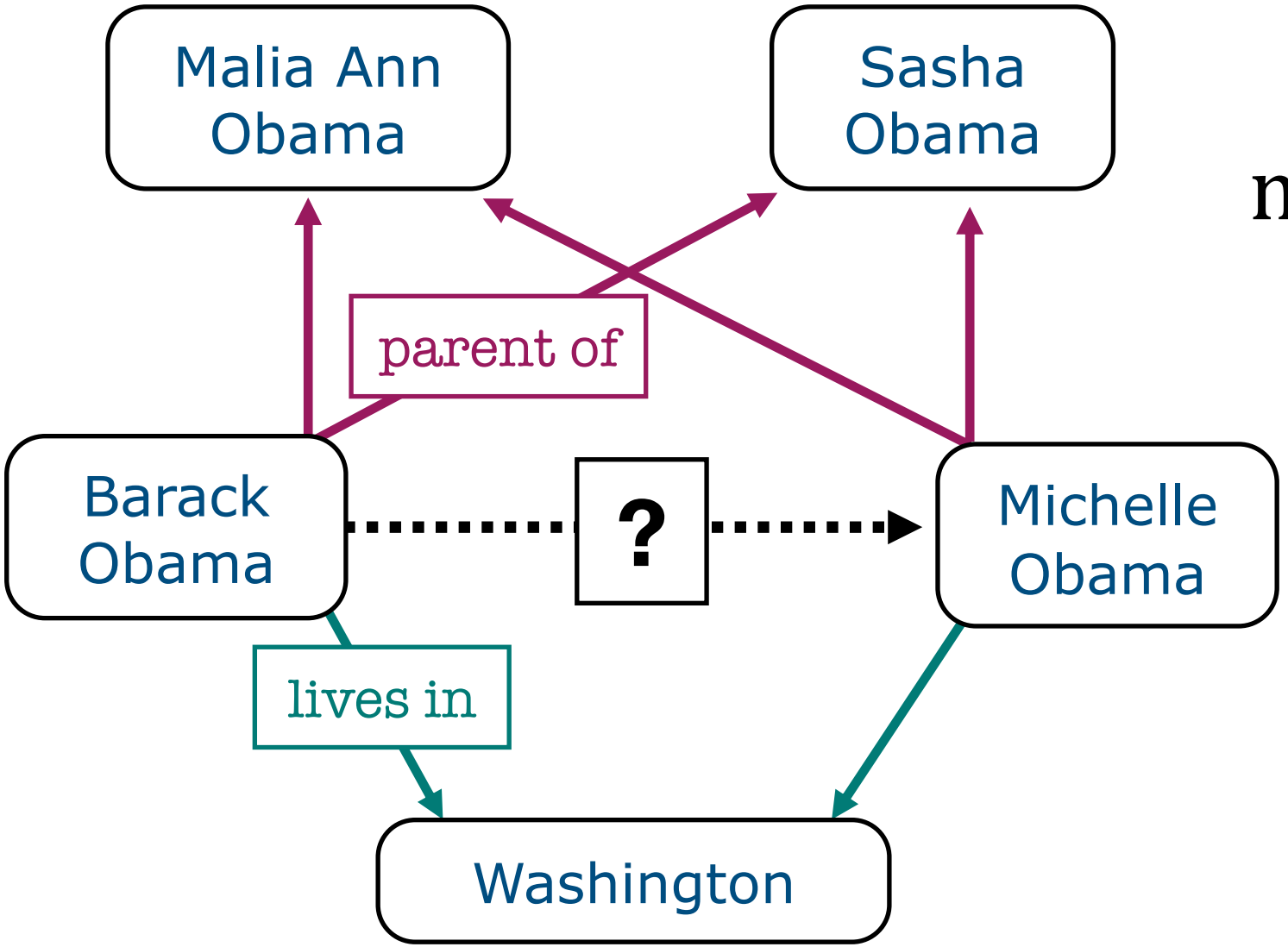


Relationships between entities *s* and *o* can be inferred from the interactions of their latent features $\mathbf{e}_s, \mathbf{e}_o$:

$$f(s,p,o) = f_p(\mathbf{e}_s, \mathbf{e}_o) \quad \begin{cases} \mathbf{e}_s, \mathbf{e}_o \in \mathbb{R}^k, \\ f_p : \mathbb{R}^k \times \mathbb{R}^k \mapsto \mathbb{R} \end{cases}$$

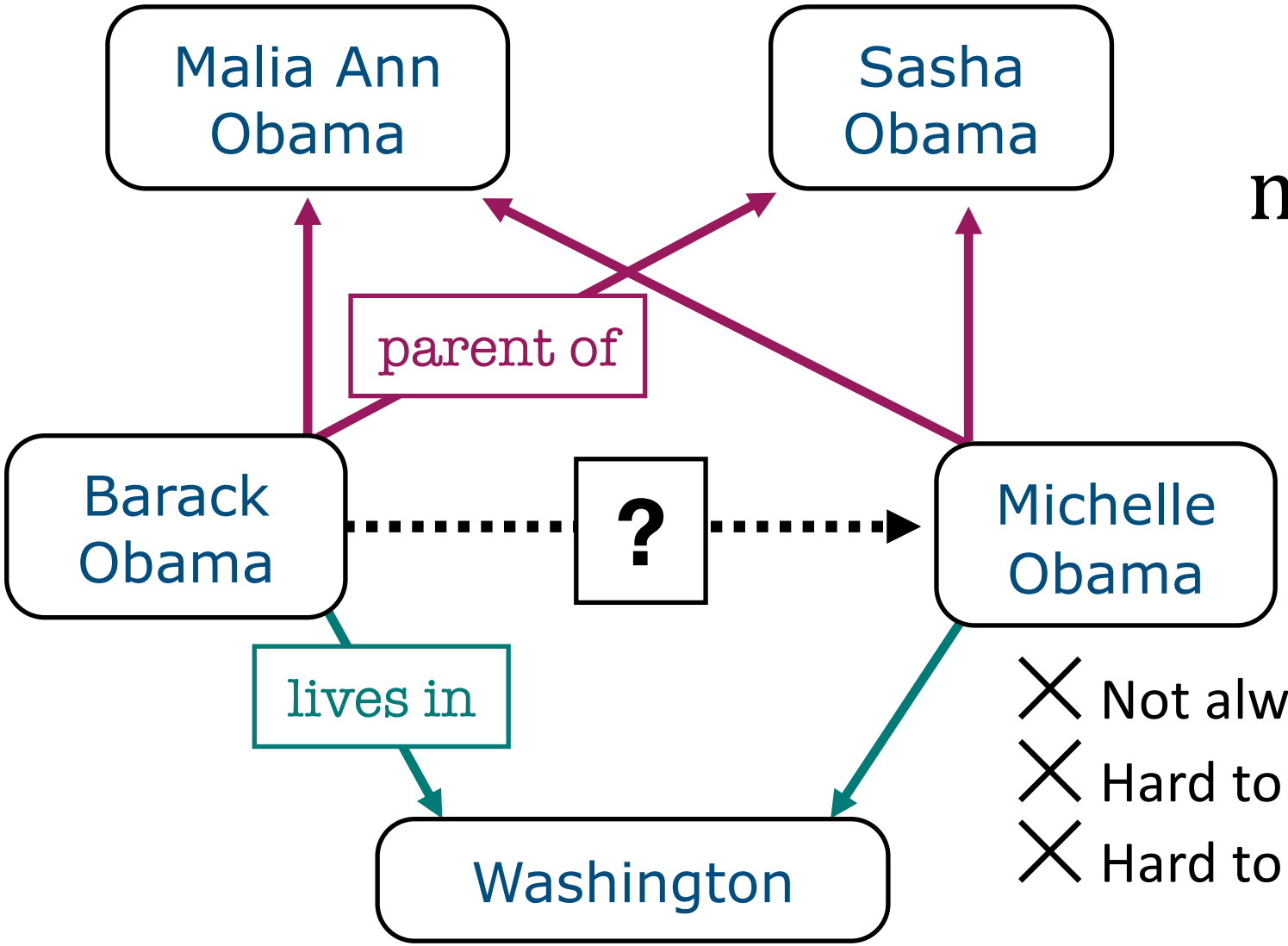
The latent features inferred by these models can be very hard to interpret.

Latent Feature Models



$$\forall X, Y, Z: \text{marriedwith}(X, Y) \Leftarrow \text{parentof}(X, Z), \text{parentof}(Y, Z)$$

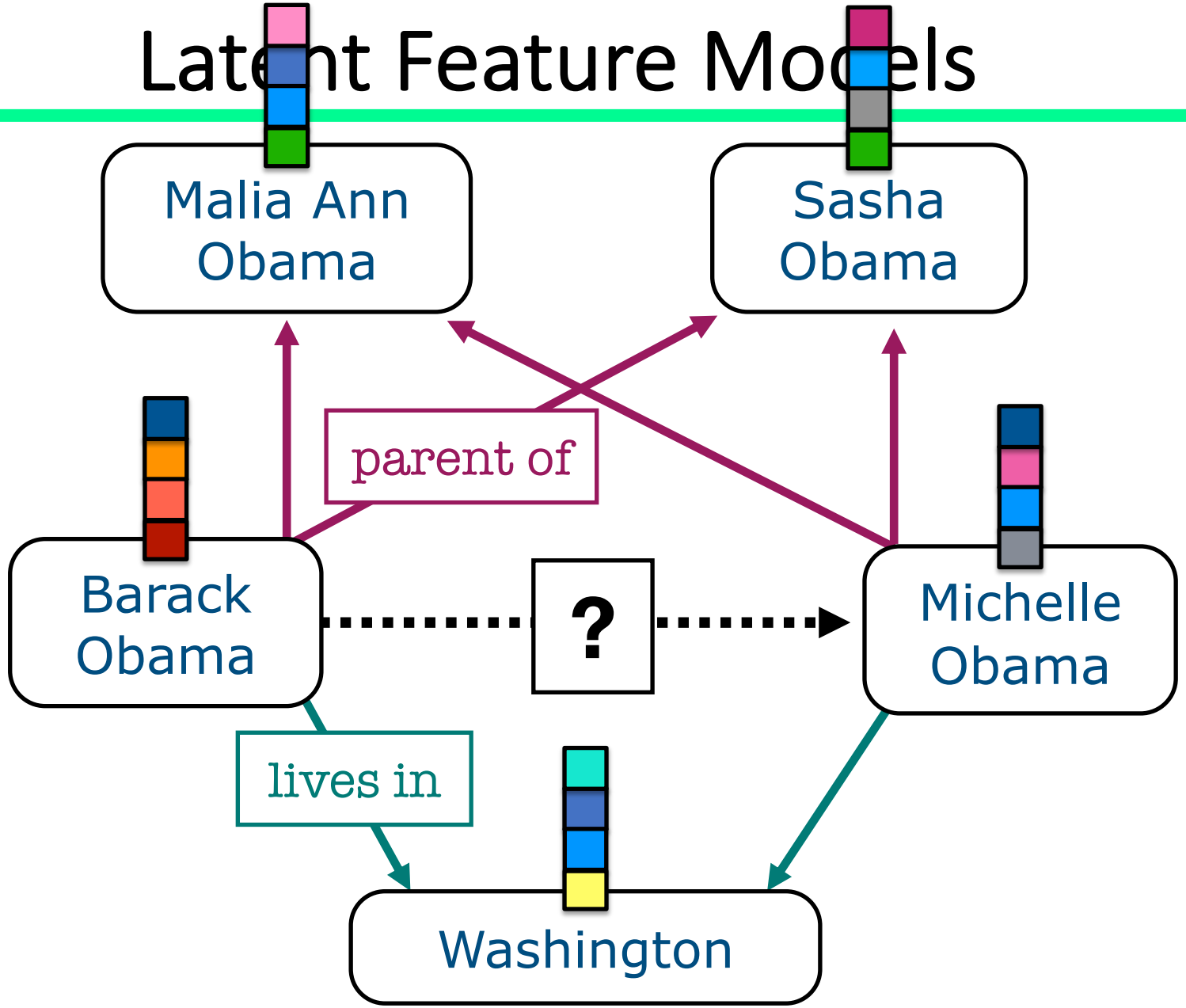
Latent Feature Models



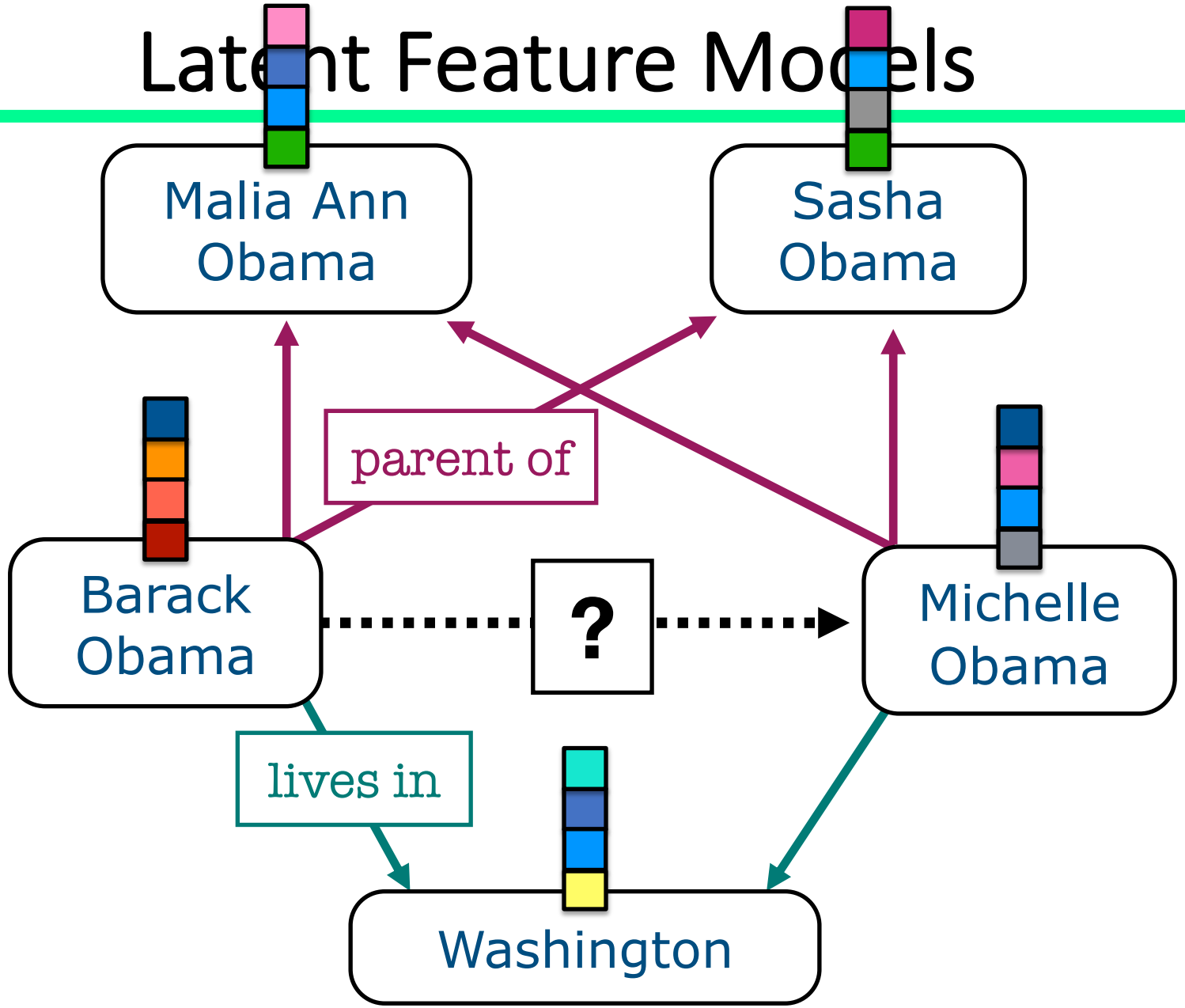
$$\forall X, Y, Z: \text{marriedwith}(X, Y) \Leftarrow \text{parentof}(X, Z), \text{parentof}(Y, Z)$$

- ✗ Not always true
- ✗ Hard to learn from data
- ✗ Hard to formalise for other modalities

Latent Feature Models



Latent Feature Models



$$P(\text{BO} \xrightarrow{\text{married}} \text{MO}) \propto$$

$$f_{\text{married}}\left(\quad , \quad \right)$$

Learning Representations

$$\mathscr{L}(\mathscr{G} \mid \Theta) = \sum_{(s,p,o) \in \mathscr{G}} \log \sigma\left(f_p(\mathbf{e}_s, \mathbf{e}_o)\right) + \sum_{(s,p,o) \notin \mathscr{G}} \log\left[1 - \sigma\left(f_p(\mathbf{e}_s, \mathbf{e}_o)\right)\right]$$

Latent Feature Models - Scoring Functions

- Relationships between entities are determined by interactions between latent features — this yields different choices for the scoring function $f_p : \mathbb{R}^k \times \mathbb{R}^k \mapsto \mathbb{R}$:

Models	Scoring Functions	Parameters
RESCAL [Nickel et al. 2011]	$\mathbf{e}_s^\top \mathbf{W}_p \mathbf{e}_o$	$\mathbf{W}_p \in \mathbb{R}^{k \times k}$
NTN [Socher et al. 2013]	$\mathbf{u}_p^\top f\left(\mathbf{e}_s \mathbf{W}_p^{[1\ldots d]} + \mathbf{V}_p \begin{bmatrix}\mathbf{e}_s \\ \mathbf{e}_o\end{bmatrix} + \mathbf{b}_p\right)$	$\mathbf{W}_p \in \mathbb{R}^{k^2 \times d}, \mathbf{V}_p \in \mathbb{R}^{2k \times d}, \mathbf{b}_p, \mathbf{u}_p \in \mathbb{R}^k$
TransE [Bordes et al. 2013]	$- \left\| \mathbf{e}_s + \mathbf{r}_p - \mathbf{e}_o \right\|_{1,2}^2$	$\mathbf{r}_p \in \mathbb{R}^k$
DistMult [Yang et al. 2014]	$\langle \mathbf{e}_s, \mathbf{r}_p, \mathbf{e}_o \rangle$	$\mathbf{r}_p \in \mathbb{R}^k$
HolE [Nickel et al. 2016]	$\mathbf{r}_p^\top \left(\mathcal{F}^{-1}\left[\overline{\mathcal{F}[\mathbf{e}_s]} \odot \mathcal{F}[\mathbf{e}_o]\right]\right)$	$\mathbf{r}_p \in \mathbb{R}^k$
ComplEx [Trouillon et al. 2016]	$\mathrm{Re}\left(\langle \mathbf{e}_s, \mathbf{r}_p, \overline{\mathbf{e}}_o \rangle\right)$	$\mathbf{r}_p \in \mathbb{C}^k$
ConvE [Dettmers et al. 2017]	$f\left(\mathrm{vec}\left(f\left([\overline{\mathbf{e}_s}; \overline{\mathbf{r}_p}] * \omega\right)\right)\mathbf{W}\right)\mathbf{e}_o$	$\mathbf{r}_p \in \mathbb{R}^k, \mathbf{W} \in \mathbb{R}^{c \times k}$

Latent Feature Models - Scoring Functions

- Relationships between entities are determined by interactions between latent features — this yields different choices for the scoring function $f_p : \mathbb{R}^k \times \mathbb{R}^k \mapsto \mathbb{R}$:

Models	Scoring Functions	Parameters
RESCAL [Nickel et al. 2011]	$\mathbf{e}_s^\top \mathbf{W}_p \mathbf{e}_o$	$\mathbf{W}_p \in \mathbb{R}^{k \times k}$
NTN [Socher et al. 2013]	$\mathbf{u}_p^\top f\left(\mathbf{e}_s \mathbf{W}_p^{[1\ldots d]} + \mathbf{V}_p \begin{bmatrix}\mathbf{e}_s \\ \mathbf{e}_o\end{bmatrix} + \mathbf{b}_p\right)$	$\mathbf{W}_p \in \mathbb{R}^{k^2 \times d}, \mathbf{V}_p \in \mathbb{R}^{2k \times d}, \mathbf{b}_p, \mathbf{u}_p \in \mathbb{R}^k$
TransE [Bordes et al. 2013]	$- \left\| \mathbf{e}_s + \mathbf{r}_p - \mathbf{e}_o \right\|_{1,2}^2$	$\mathbf{r}_p \in \mathbb{R}^k$
DistMult [Yang et al. 2014]	$\langle \mathbf{e}_s, \mathbf{r}_p, \mathbf{e}_o \rangle$	$\mathbf{r}_p \in \mathbb{R}^k$
HolE [Nickel et al. 2016]	$\mathbf{r}_p^\top \left(\mathcal{F}^{-1}\left[\overline{\mathcal{F}[\mathbf{e}_s]} \odot \mathcal{F}[\mathbf{e}_o]\right]\right)$	$\mathbf{r}_p \in \mathbb{R}^k$
ComplEx [Trouillon et al. 2016]	$\mathrm{Re}\left(\langle \mathbf{e}_s, \mathbf{r}_p, \overline{\mathbf{e}}_o \rangle\right)$	$\mathbf{r}_p \in \mathbb{C}^k$
ConvE [Dettmers et al. 2017]	$f\left(\mathrm{vec}\left(f\left([\overline{\mathbf{e}_s}; \overline{\mathbf{r}_p}] * \omega\right)\right)\mathbf{W}\right)\mathbf{e}_o$	$\mathbf{r}_p \in \mathbb{R}^k, \mathbf{W} \in \mathbb{R}^{c \times k}$

Latent Feature Models - Scoring Functions

- Relationships between entities are determined by interactions between latent features — this yields different choices for the scoring function $f_p : \mathbb{R}^k \times \mathbb{R}^k \mapsto \mathbb{R}$:

Models	Scoring Functions	Parameters
RESCAL [Nickel et al. 2011]	$\mathbf{e}_s^\top \mathbf{W}_p \mathbf{e}_o$	$\mathbf{W}_p \in \mathbb{R}^{k \times k}$
NTN [Socher et al. 2013]	$\mathbf{u}_p^\top f\left(\mathbf{e}_s \mathbf{W}_p^{[1\ldots d]} + \mathbf{V}_p \begin{bmatrix}\mathbf{e}_s \\ \mathbf{e}_o\end{bmatrix} + \mathbf{b}_p\right)$	$\mathbf{W}_p \in \mathbb{R}^{k^2 \times d}, \mathbf{V}_p \in \mathbb{R}^{2k \times d}, \mathbf{b}_p, \mathbf{u}_p \in \mathbb{R}^k$
TransE [Bordes et al. 2013]	$- \left\| \mathbf{e}_s + \mathbf{r}_p - \mathbf{e}_o \right\|_{1,2}^2$	$\mathbf{r}_p \in \mathbb{R}^k$
DistMult [Yang et al. 2014]	$\langle \mathbf{e}_s, \mathbf{r}_p, \mathbf{e}_o \rangle$	$\mathbf{r}_p \in \mathbb{R}^k$
HolE [Nickel et al. 2016]	$\mathbf{r}_p^\top \left(\mathcal{F}^{-1}\left[\overline{\mathcal{F}[\mathbf{e}_s]} \odot \mathcal{F}[\mathbf{e}_o]\right]\right)$	$\mathbf{r}_p \in \mathbb{R}^k$
ComplEx [Trouillon et al. 2016]	$\mathrm{Re}\left(\langle \mathbf{e}_s, \mathbf{r}_p, \overline{\mathbf{e}}_o \rangle\right)$	$\mathbf{r}_p \in \mathbb{C}^k$
ConvE [Dettmers et al. 2017]	$f\left(\mathrm{vec}\left(f\left([\overline{\mathbf{e}_s}; \overline{\mathbf{r}_p}] * \omega\right)\right)\mathbf{W}\right)\mathbf{e}_o$	$\mathbf{r}_p \in \mathbb{R}^k, \mathbf{W} \in \mathbb{R}^{c \times k}$

Latent Feature Models - Scoring Functions

- Relationships between entities are determined by interactions between latent features — this yields different choices for the scoring function $f_p : \mathbb{R}^k \times \mathbb{R}^k \mapsto \mathbb{R}$:

Models	Scoring Functions	Parameters
RESCAL [Nickel et al. 2011]	$\mathbf{e}_s^\top \mathbf{W}_p \mathbf{e}_o$	$\mathbf{W}_p \in \mathbb{R}^{k \times k}$
NTN [Socher et al. 2013]	$\mathbf{u}_p^\top f\left(\mathbf{e}_s \mathbf{W}_p^{[1\ldots d]} + \mathbf{V}_p \begin{bmatrix}\mathbf{e}_s \\ \mathbf{e}_o\end{bmatrix} + \mathbf{b}_p\right)$	$\mathbf{W}_p \in \mathbb{R}^{k^2 \times d}, \mathbf{V}_p \in \mathbb{R}^{2k \times d}, \mathbf{b}_p, \mathbf{u}_p \in \mathbb{R}^k$
TransE [Bordes et al. 2013]	$- \left\| \mathbf{e}_s + \mathbf{r}_p - \mathbf{e}_o \right\|_{1,2}^2$	$\mathbf{r}_p \in \mathbb{R}^k$
DistMult [Yang et al. 2014]	$\langle \mathbf{e}_s, \mathbf{r}_p, \mathbf{e}_o \rangle$	$\mathbf{r}_p \in \mathbb{R}^k$
HolE [Nickel et al. 2016]	$\mathbf{r}_p^\top \left(\mathcal{F}^{-1}\left[\overline{\mathcal{F}[\mathbf{e}_s]} \odot \mathcal{F}[\mathbf{e}_o]\right]\right)$	$\mathbf{r}_p \in \mathbb{R}^k$
ComplEx [Trouillon et al. 2016]	$\mathrm{Re}\left(\langle \mathbf{e}_s, \mathbf{r}_p, \overline{\mathbf{e}}_o \rangle\right)$	$\mathbf{r}_p \in \mathbb{C}^k$
ConvE [Dettmers et al. 2017]	$f\left(\mathrm{vec}\left(f\left([\overline{\mathbf{e}_s}; \overline{\mathbf{r}_p}] * \omega\right)\right)\mathbf{W}\right)\mathbf{e}_o$	$\mathbf{r}_p \in \mathbb{R}^k, \mathbf{W} \in \mathbb{R}^{c \times k}$

Latent Feature Models - Predictive Accuracy

Evaluation Metrics — Area Under the Precision-Recall Curve (AUC-PR), Mean Reciprocal Rank (MRR), Hits@k. In MRR and Hits@k, for each test triple:

- Modify its subject with all the entities in the Knowledge Graph,
- Score all the triple variants, and *compute the rank* of the original test triple,
- Repeat for the object.

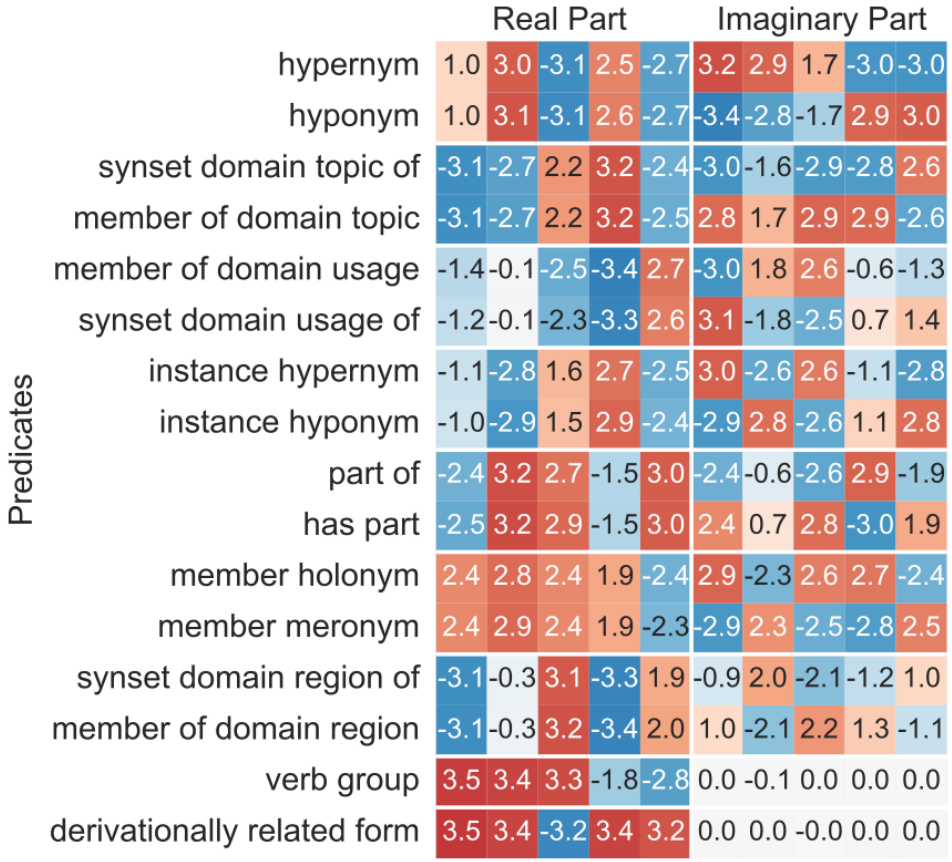
From [Lacroix et al. ICML 2018]

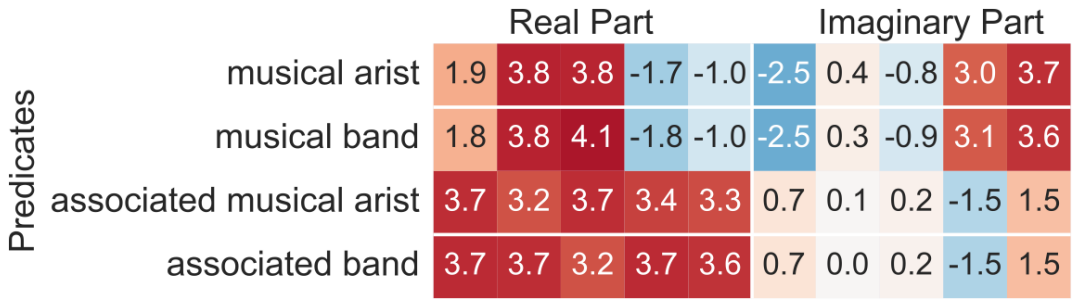
$$MRR = \frac{1}{|\mathcal{T}|}\sum_{i=1}^{|\mathcal{T}|}\frac{1}{rank_i}, HITS@k = \frac{|\{rank_i \leq 10\}|}{|\mathcal{T}|}$$

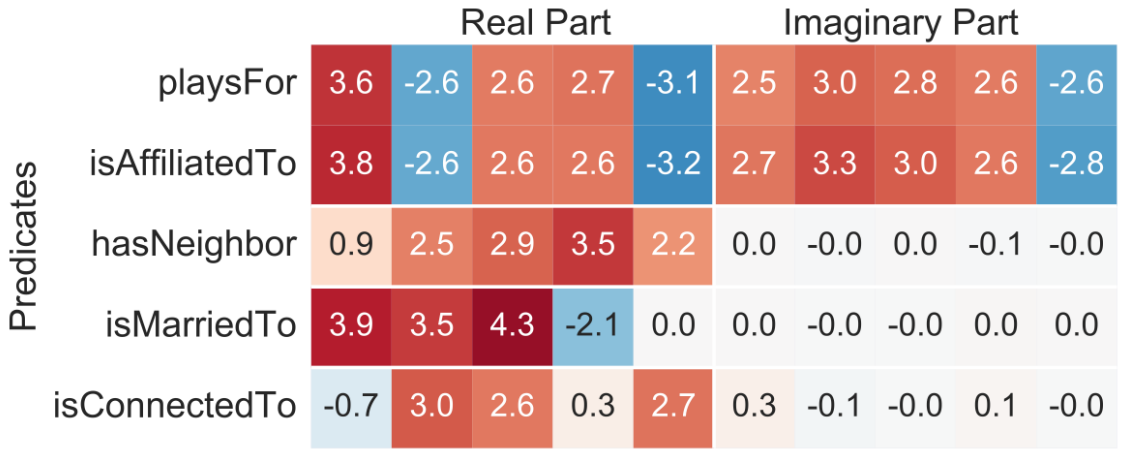
	Model	WN18		WN18RR		FB15K		FB15K-237		YAGO3-10	
		MRR	H@10	MRR	H@10	MRR	H@10	MRR	H@10	MRR	H@10
Reciprocal	CP-FRO	**0.95**	0.95	0.46	0.48	**0.86**	0.91	0.34	0.51	0.54	0.68
	CP-N3	**0.95**	0.96	0.47	0.54	**0.86**	0.91	0.36	0.54	0.57	**0.71**
	ComplEx-FRO	**0.95**	0.96	0.47	0.54	**0.86**	0.91	0.35	0.53	0.57	**0.71**
	ComplEx-N3	**0.95**	0.96	**0.48**	**0.57**	**0.86**	0.91	**0.37**	**0.56**	**0.58**	**0.71**

Latent Feature Models - Interpreting the Embeddings

Learned relation embeddings — using *ComplEx* with a *pairwise margin-based loss* — for WordNet (left), DBpedia, and YAGO (right) [Minervini et al. ECML 2017]



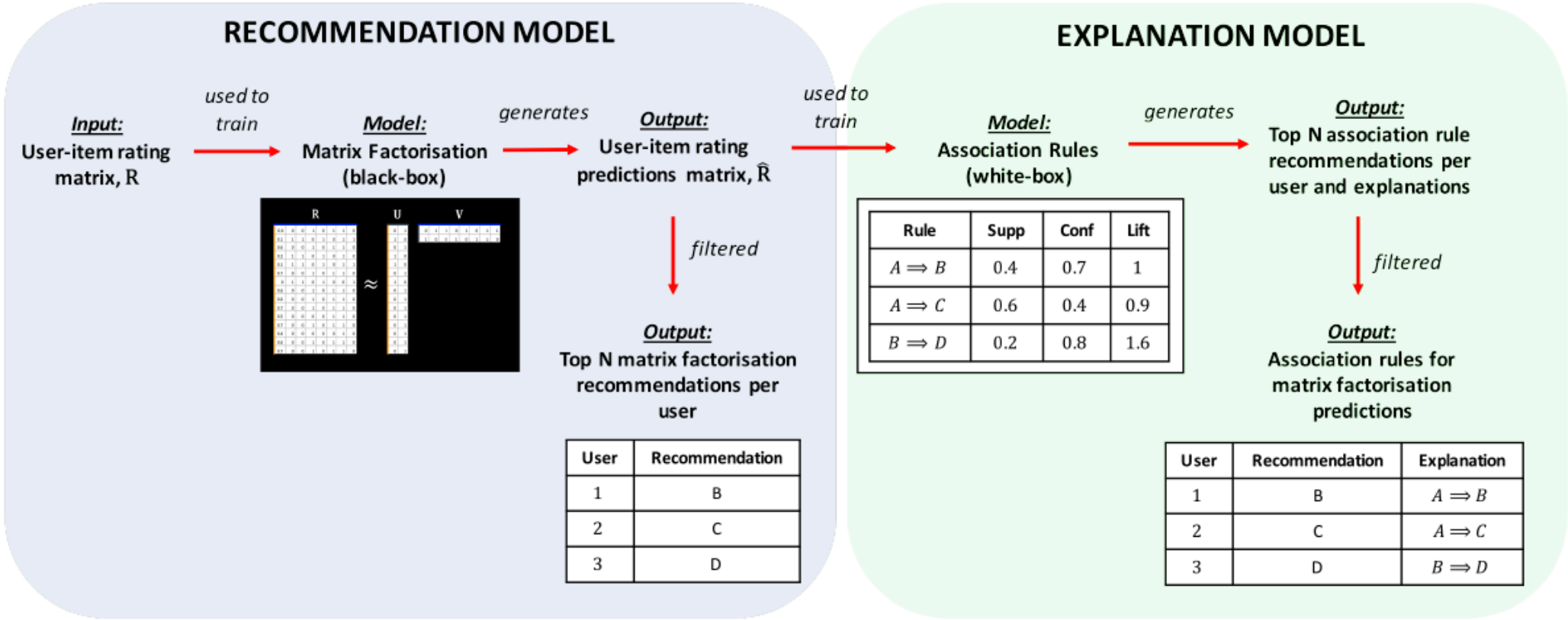




Latent Feature Models - Post Hoc Interpretability

[Carmona et al. 2015, Peake et al. KDD 2018, Gusmão et al. 2018]

- Generate an explanation model by training Bayesian Networks or Association Rules on the output of a Latent Feature Model.



Combining Observable and Latent Feature Models

- Additive Relational Effects (ARE) [Nickel et al. NeurIPS 2014] — combines Observable and Latent Features in a single linear model:

$$f_{spo}^{ARE} = \mathbf{w}_{LFM,p}^{\top} \Theta_{LFM,so} + \mathbf{w}_{OBS,p}^{\top} \Theta_{PRA,so}$$

- Knowledge Vault [Dong et al. KDD 2014] — combines the prediction of Observable and Latent Feature Models via stacking:

$$f_{spo}^{KV} = f_{FUSION}\left(f_{spo}^{OFM}, f_{spo}^{LFM}\right)$$

- Adversarial Sets [Minervini et al. UAI 2017] — incorporate observable features, in the form of First-Order Logic Rules R, in Latent Feature Models:

$$\mathcal{L}(\Theta \mid R) = \mathcal{L}_{LFM}(\Theta) + \max_{\mathcal{S} \subseteq \mathcal{P}(\mathcal{E})} \mathcal{L}_{RULE}(\Theta, R)$$

Incorporating Rules via Adversarial Training

Idea — adversarial training process where, iteratively:

- An adversary searches for inputs where the model violates constraints

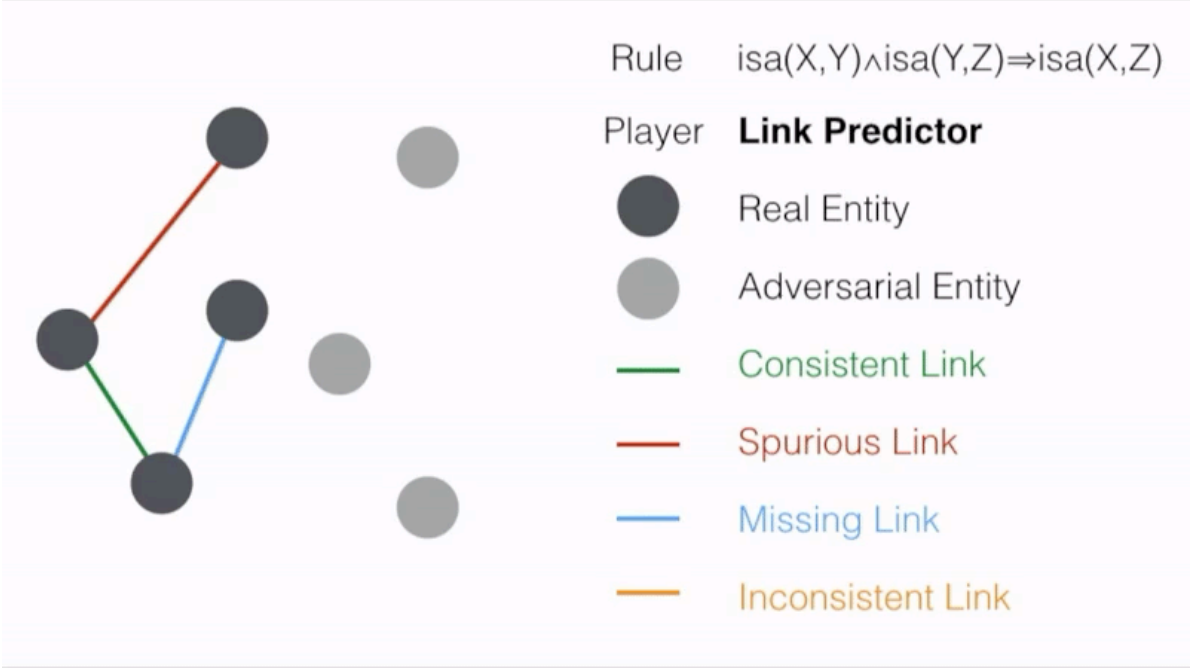
$$e.g. x, y, z \; such \; that$$

$$\text{isa}(x, y) \wedge \text{isa}(y, z) \wedge \neg \text{isa}(x, z)$$

Incorporating Rules via Adversarial Training

Idea — adversarial training process where, iteratively:

- An adversary searches for inputs where the model violates constraints
- The model is regularised to correct such violations.



Incorporating Rules via Adversarial Training

Idea — adversarial training process where, iteratively:

- An adversary searches for inputs where the model violates constraints
- The model is regularised to correct such violations.

$$\min_{\Theta} \mathscr{L}_{\text{data}}(D \mid \Theta) + \lambda \max_{S} \mathscr{L}_{\text{violation}}(S, D \mid \Theta)$$

Incorporating Rules via Adversarial Training

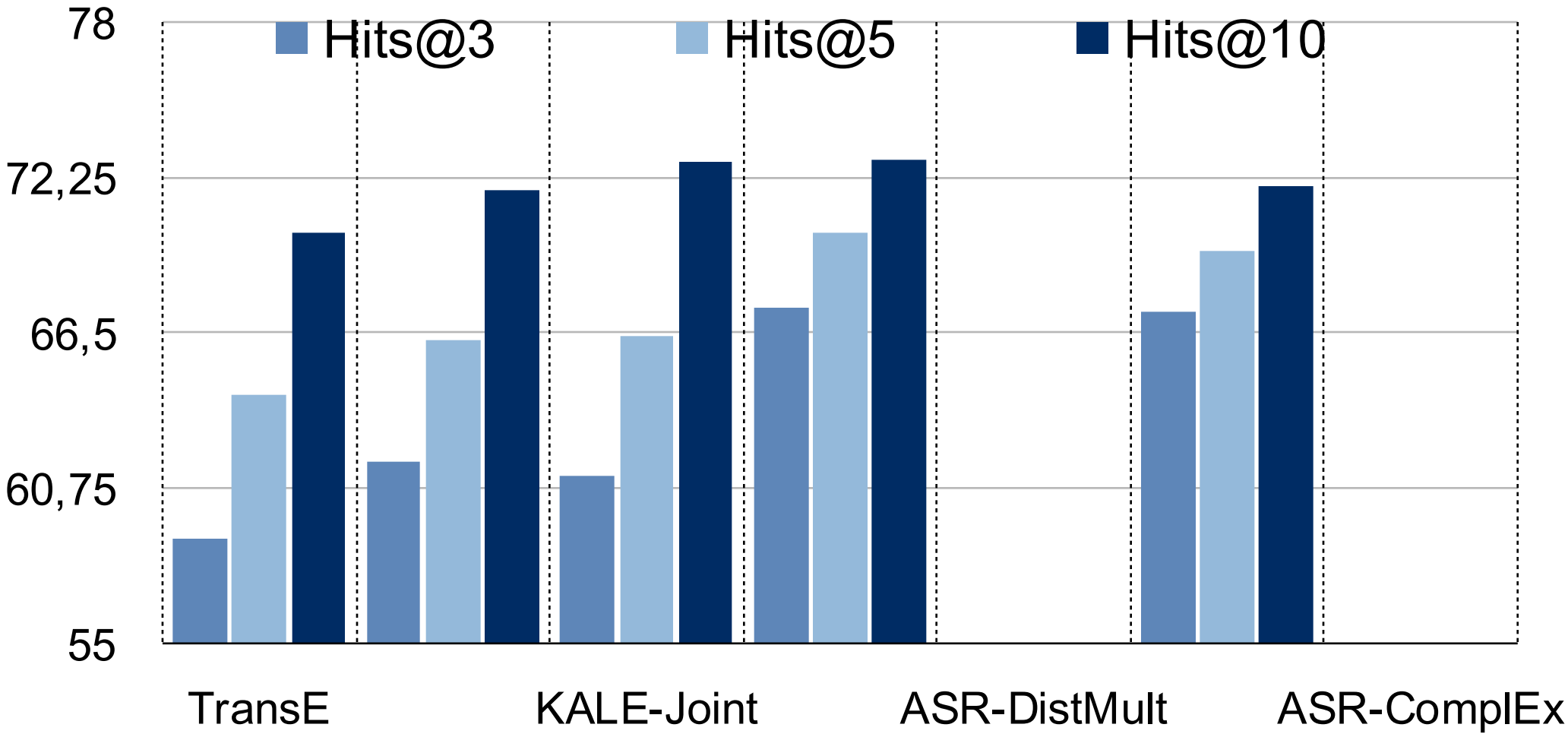
Idea — adversarial training process where, iteratively:

- An adversary searches for inputs where the model violates constraints
- The model is regularised to correct such violations.

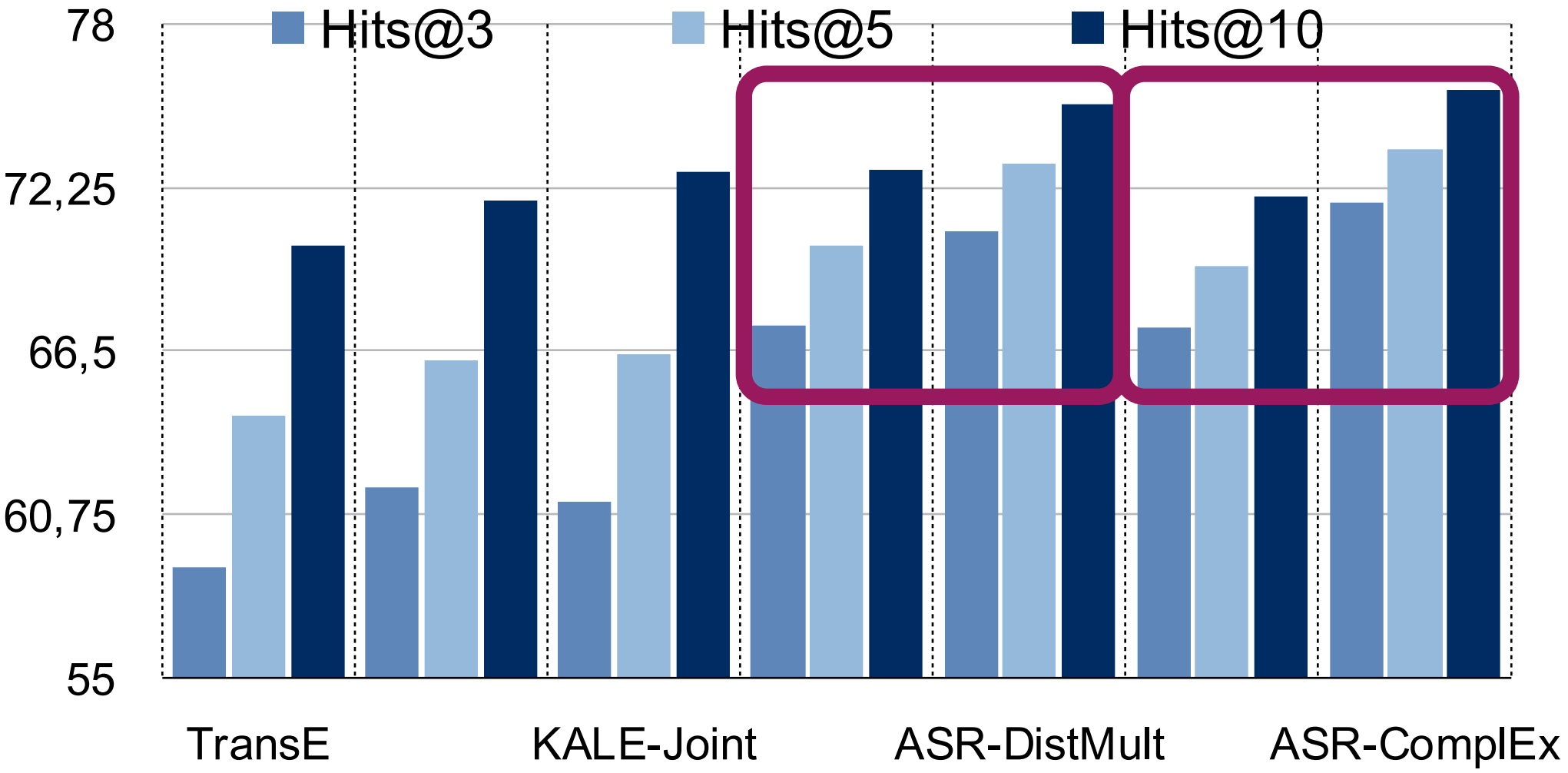
$$\min_{\Theta} \mathscr{L}_{\text{data}}(D \mid \Theta) + \lambda \max_{S} \mathscr{L}_{\text{violation}}(S, D \mid \Theta)$$

- Inputs S can be either input space or embedding space
- In most interesting cases, max has closed form solutions
- Constraints are guaranteed to hold everywhere in embedding space.

Incorporating Rules via Adversarial Training



Incorporating Rules via Adversarial Training



End-to-End Differentiable Reasoning

We can combine *neural networks* and *symbolic models* by re-implementing classic reasoning algorithms using end-to-end differentiable (neural) architectures:

Differentiable Architectures

- Can generalise from high-dimensional, noisy, ambiguous inputs (*e.g.* sensory)
- Not interpretable
- Hard to incorporate knowledge
- Propositional fixation [McCarthy, 1988]

Logic Reasoning Based Models

- Can learn from small data
- Issues with high-dimensional, noisy, ambiguous inputs (*e.g.* images)
- Easy to *interpret*, and can provide *explanations* in the form of reasoning steps used to derive a conclusion

Reasoning in a Nutshell — Forward Chaining

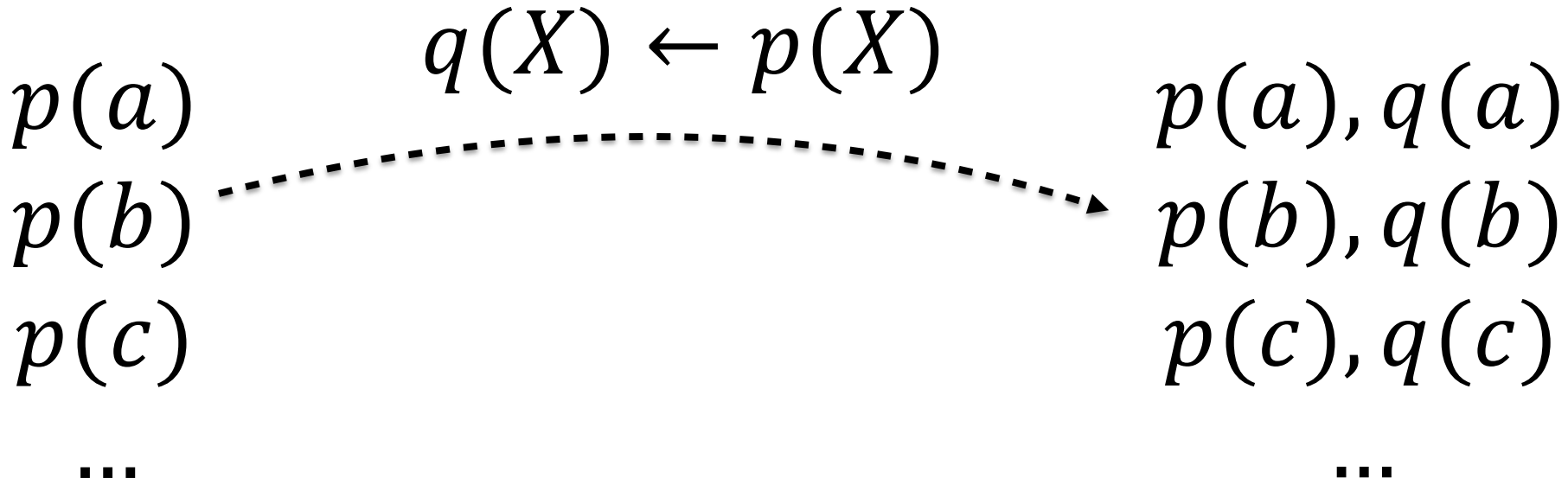
- **Forward Chaining** — start with a list of facts, and work forward from the antecedent P to the consequent Q iteratively.

$$q(X) \leftarrow p(X)$$

$$p(a)$$
$$p(b)$$
$$p(c)$$
$$\ldots$$

Reasoning in a Nutshell — Forward Chaining

- **Forward Chaining** — start with a list of facts, and work forward from the antecedent P to the consequent Q iteratively.



Reasoning in a Nutshell — Backward Chaining

- **Backward Chaining** — start with a list of goals, and work backwards from the consequent Q to the antecedent P to see if any data supports any of the consequents.

$$q(X) \leftarrow p(X)$$

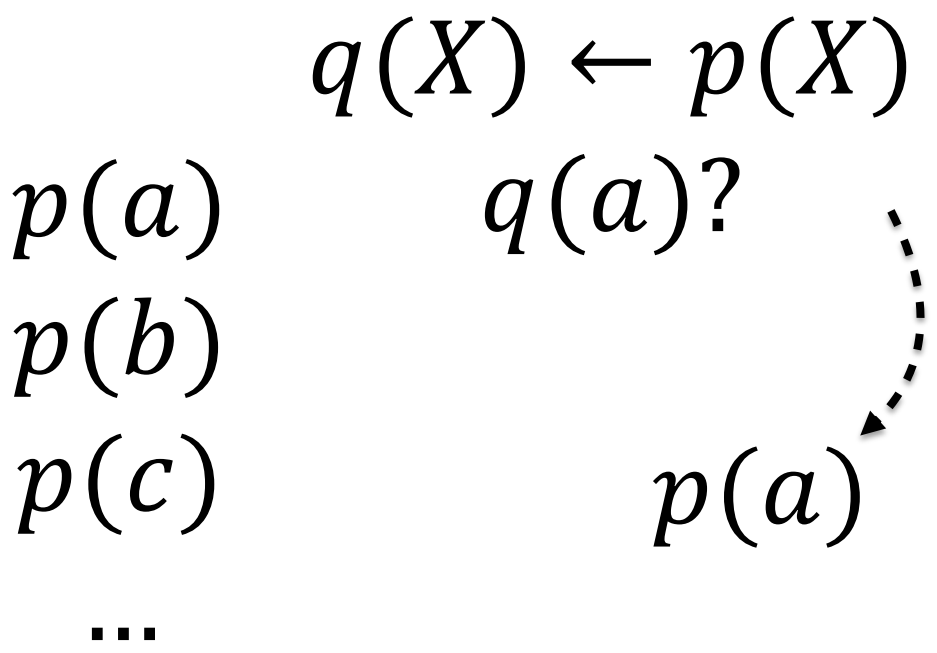
$$q(a)?$$

$$p(a)$$
$$p(b)$$
$$p(c)$$
$$\ldots$$

You can see backward chaining as a *query reformulation strategy.*

Reasoning in a Nutshell — Backward Chaining

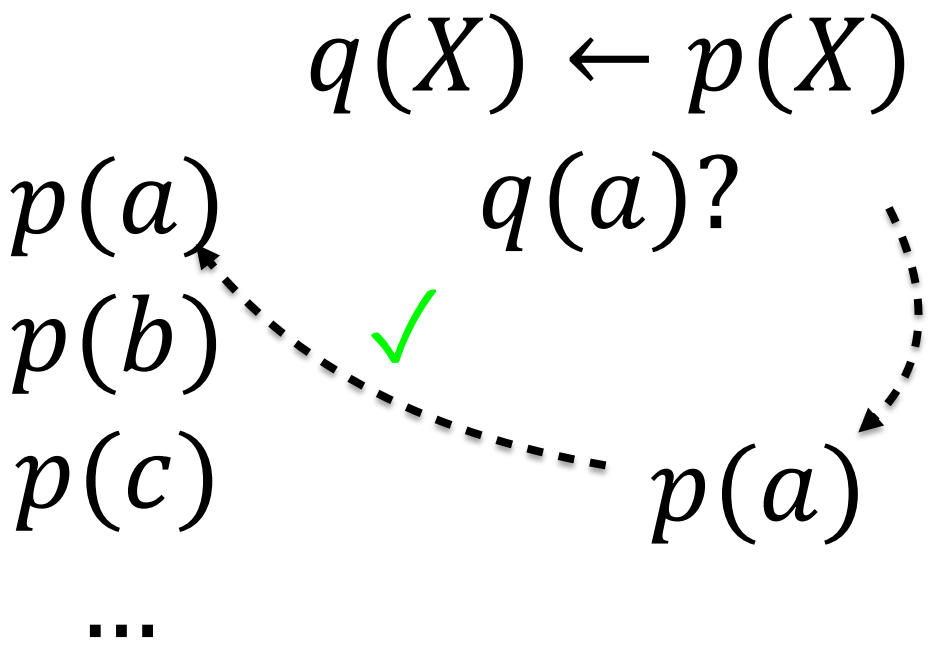
- **Backward Chaining** — start with a list of goals, and work backwards from the consequent Q to the antecedent P to see if any data supports any of the consequents.



You can see backward chaining as a *query reformulation strategy.*

Reasoning in a Nutshell — Backward Chaining

- **Backward Chaining** — start with a list of goals, and work backwards from the consequent Q to the antecedent P to see if any data supports any of the consequents.

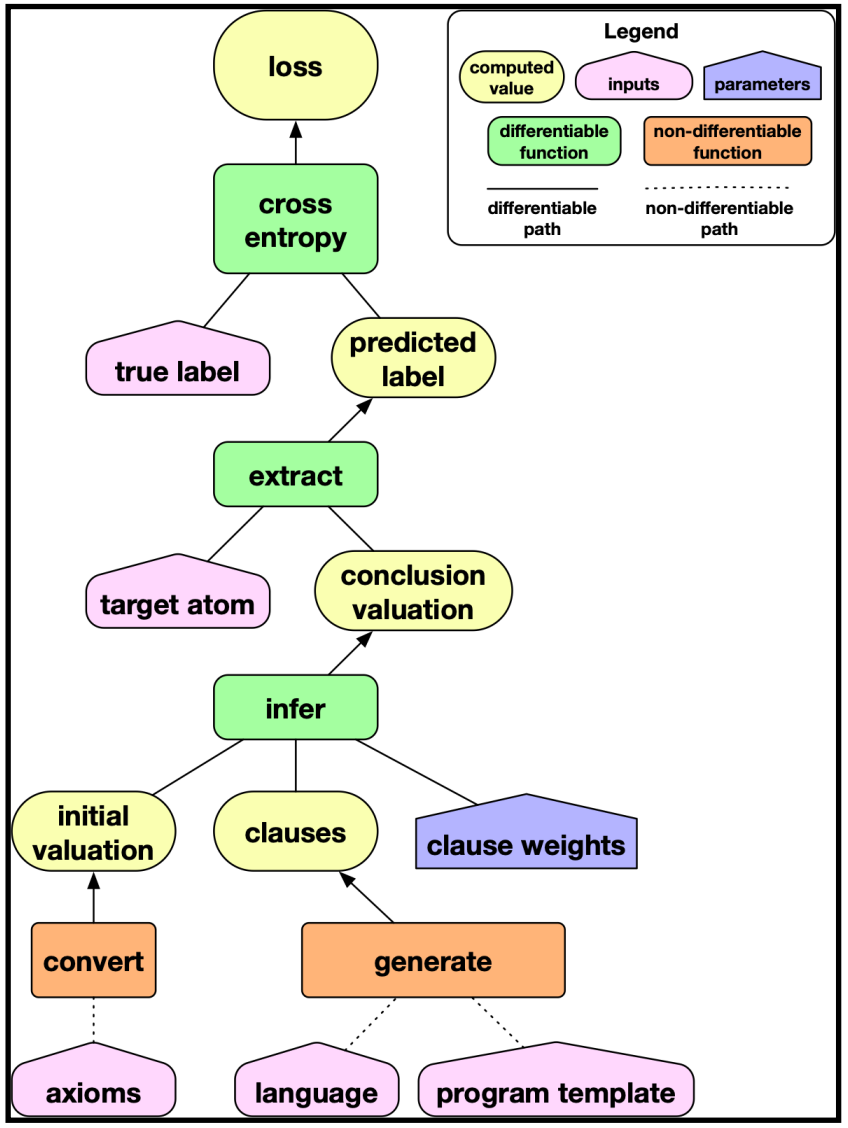


You can see backward chaining as a *query reformulation strategy.*

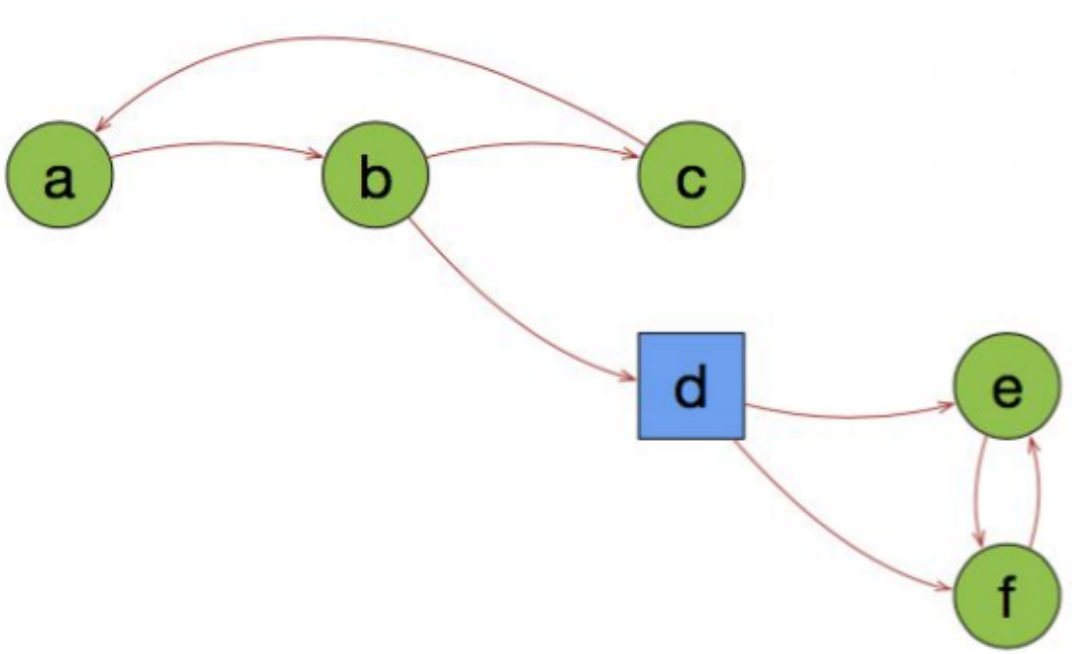
Differentiable Forward Chaining - ∂ILP [Evans et al. JAIR 2018]

∂ILP uses a *differentiable model* of forward chaining inference:

- Weights of the network represent a probability distribution over clauses
- A valuation is a vector with values in [0, 1] representing how likely it is that each of the ground atoms is true
- Forward chaining is implemented by a differentiable function that, given a valuation vector, produces another by applying rules to it.
- If conclusions do not match the desired ones, the error is back-propagated to the weights.
- We can extract a readable program.



Differentiable Forward Chaining - ∂ILP [Evans et al. JAIR 2018]



$$\text{cycle}(X) \leftarrow \text{pred}(X, X)$$
$$\text{pred}(X, Y) \leftarrow \text{edge}(X, Y)$$
$$\text{pred}(X, Y) \leftarrow \text{edge}(X, Z), \text{pred}(Z, Y)$$

Differentiable Forward Chaining - ∂ILP [Evans et al. JAIR 2018]

$$
\begin{aligned}
1 &\mapsto 1 \\
2 &\mapsto 2 \\
3 &\mapsto Fizz \\
4 &\mapsto 4 \\
5 &\mapsto Buzz \\
6 &\mapsto Fizz \\
7 &\mapsto 7 \\
8 &\mapsto 8 \\
9 &\mapsto Fizz \\
10 &\mapsto Buzz
\end{aligned}
$$

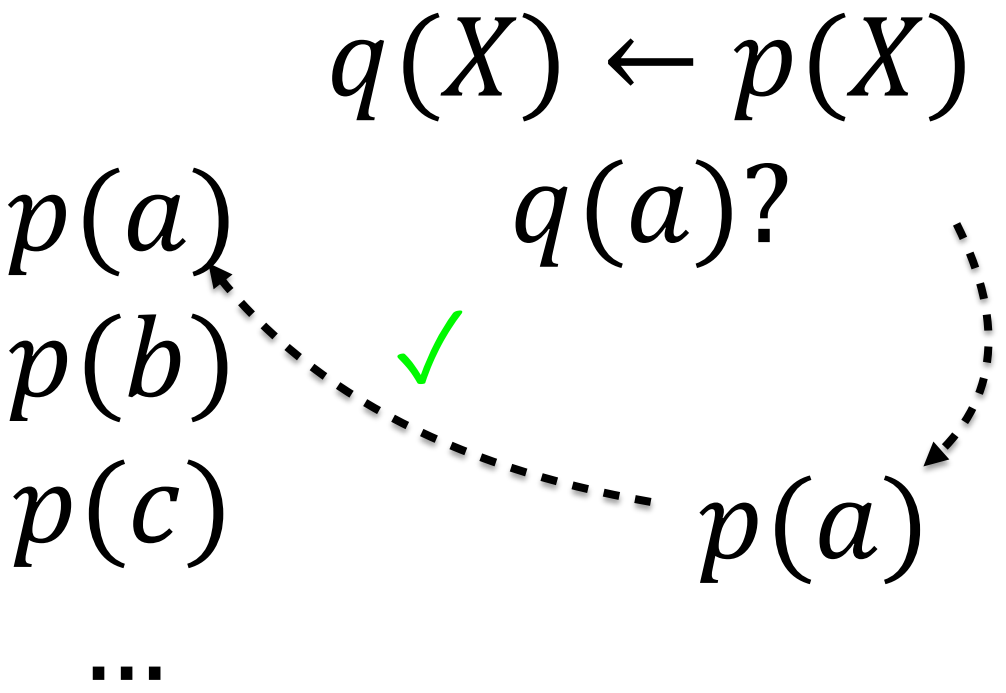
Differentiable Forward Chaining - ∂ILP [Evans et al. JAIR 2018]

$$1 \mapsto 1$$
$$2 \mapsto 2$$
$$3 \mapsto Fizz$$
$$4 \mapsto 4$$
$$5 \mapsto Buzz$$
$$6 \mapsto Fizz$$
$$7 \mapsto 7$$
$$8 \mapsto 8$$
$$9 \mapsto Fizz$$
$$10 \mapsto Buzz$$

$$\text{fizz}(X) \leftarrow \text{zero}(X)$$
$$\text{fizz}(X) \leftarrow \text{fizz}(Y), \text{pred1}(Y, X)$$
$$\text{pred1}(X, Y) \leftarrow \text{succ}(X, Z), \text{pred2}(Z, Y)$$
$$\text{pred2}(X, Y) \leftarrow \text{succ}(X, Z), \text{succ}(Z, Y)$$

Backward Chaining — Differentiable Proving

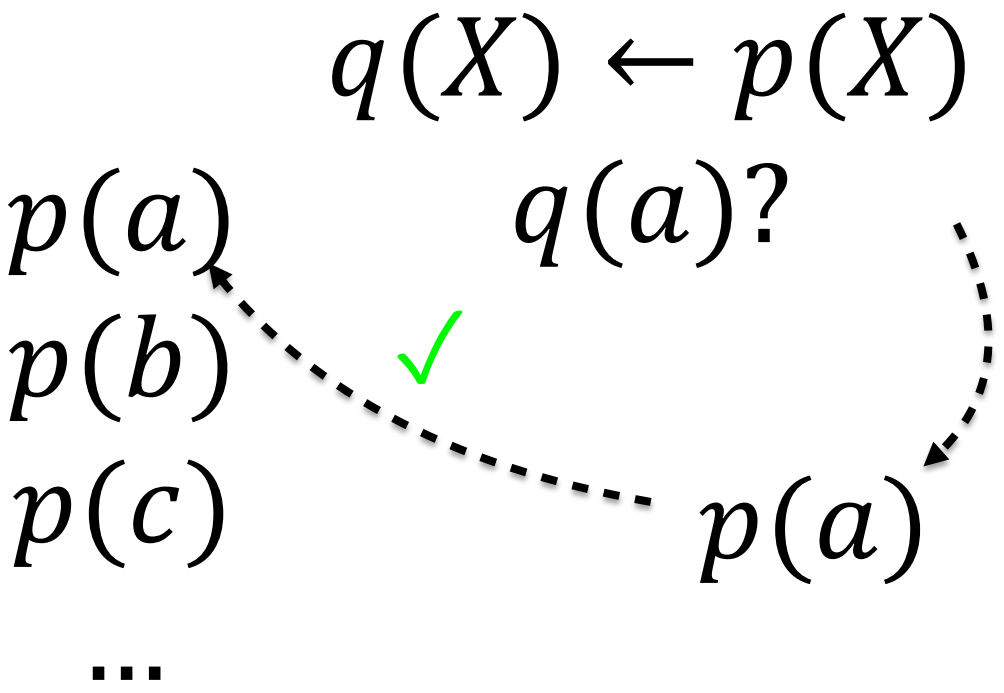
Backward Chaining



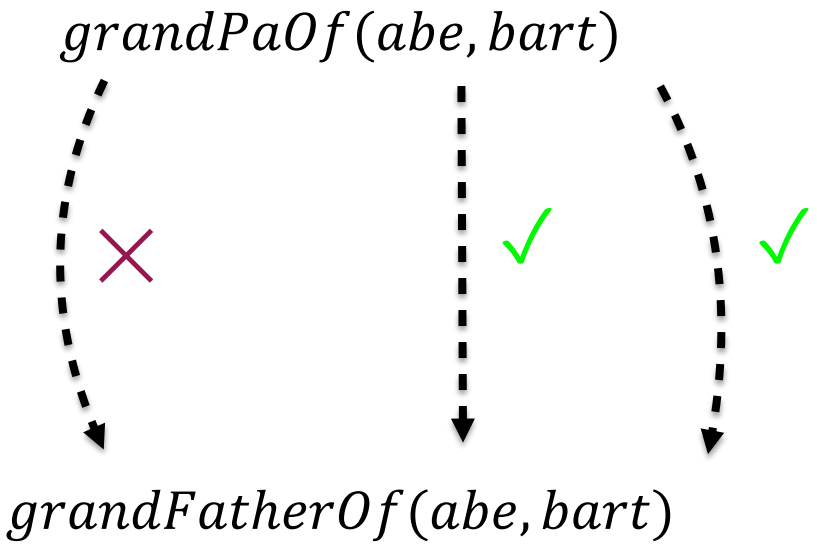
[Rocktäschel et al. 2017, Minervini et al. 2018, Welbl et al. 2019]

Backward Chaining — Differentiable Proving

Backward Chaining

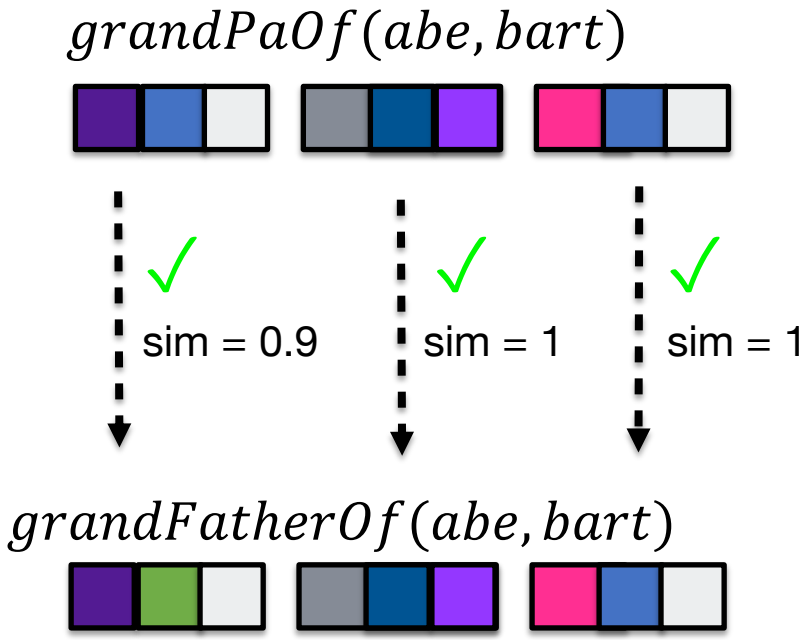


BUT there's a problem..



[Rocktäschel et al. 2017, Minervini et al. 2018, Welbl et al. 2019]

Backward Chaining — Differentiable Proving



[Rocktäschel et al. 2017, Minervini et al. 2018, Welbl et al. 2019]

Backward Chaining — Differentiable Proving

Knowledge Base:

$$fatherOf(abe, homer)$$
$$parentOf(homer, bart)$$
$$grandFatherOf(X, Y) \Leftarrow fatherOf(X, Z), parentOf(Z, Y).$$

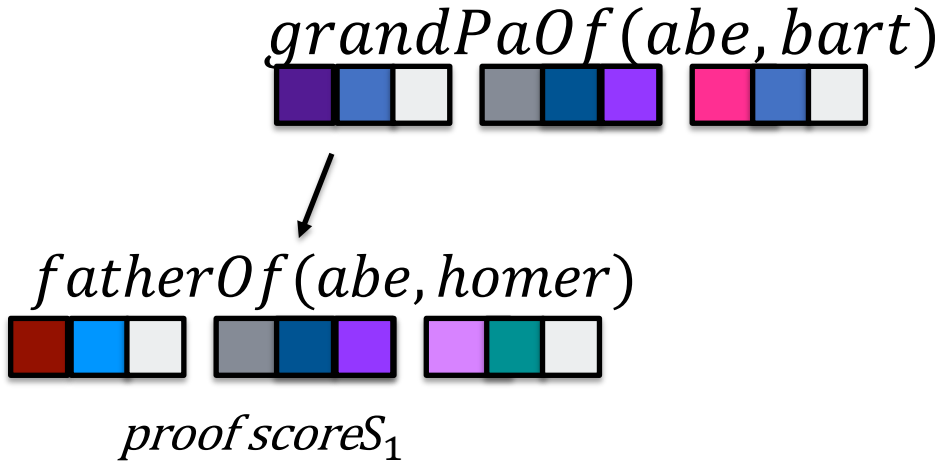
$$grandPaOf(abe, bart)$$

[Rocktäschel et al. 2017, Minervini et al. 2018, Welbl et al. 2019]

Backward Chaining — Differentiable Proving

Knowledge Base:

$fatherOf(abe, homer)$
$parentOf(homer, bart)$
$grandFatherOf(X, Y) \Leftarrow$
$fatherOf(X, Z),$
$parentOf(Z, Y).$

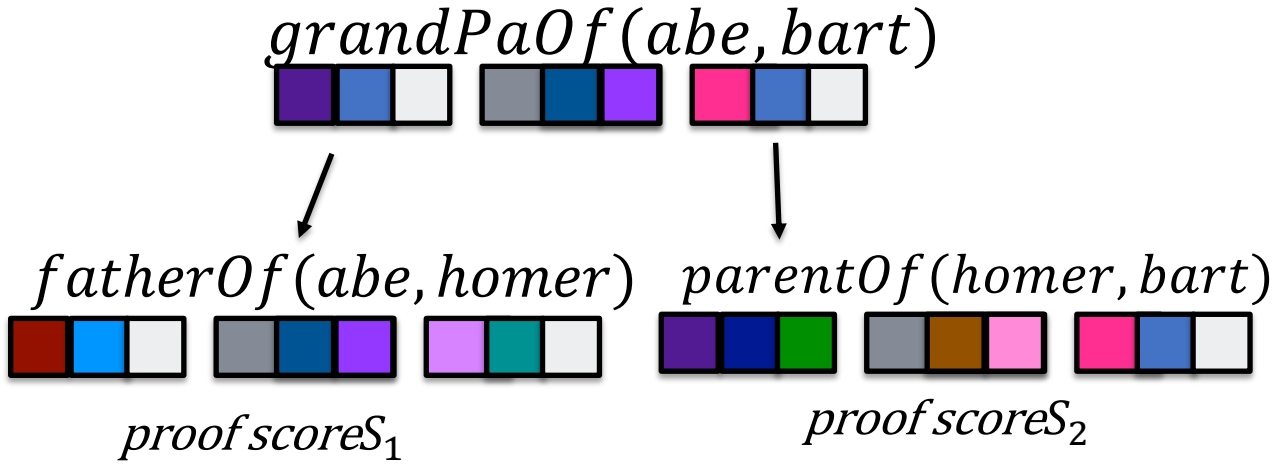


[Rocktäschel et al. 2017, Minervini et al. 2018, Welbl et al. 2019]

Backward Chaining — Differentiable Proving

Knowledge Base:

$fatherOf(abe, homer)$
$parentOf(homer, bart)$
$grandFatherOf(X, Y) \Leftarrow$
$fatherOf(X, Z),$
$parentOf(Z, Y).$

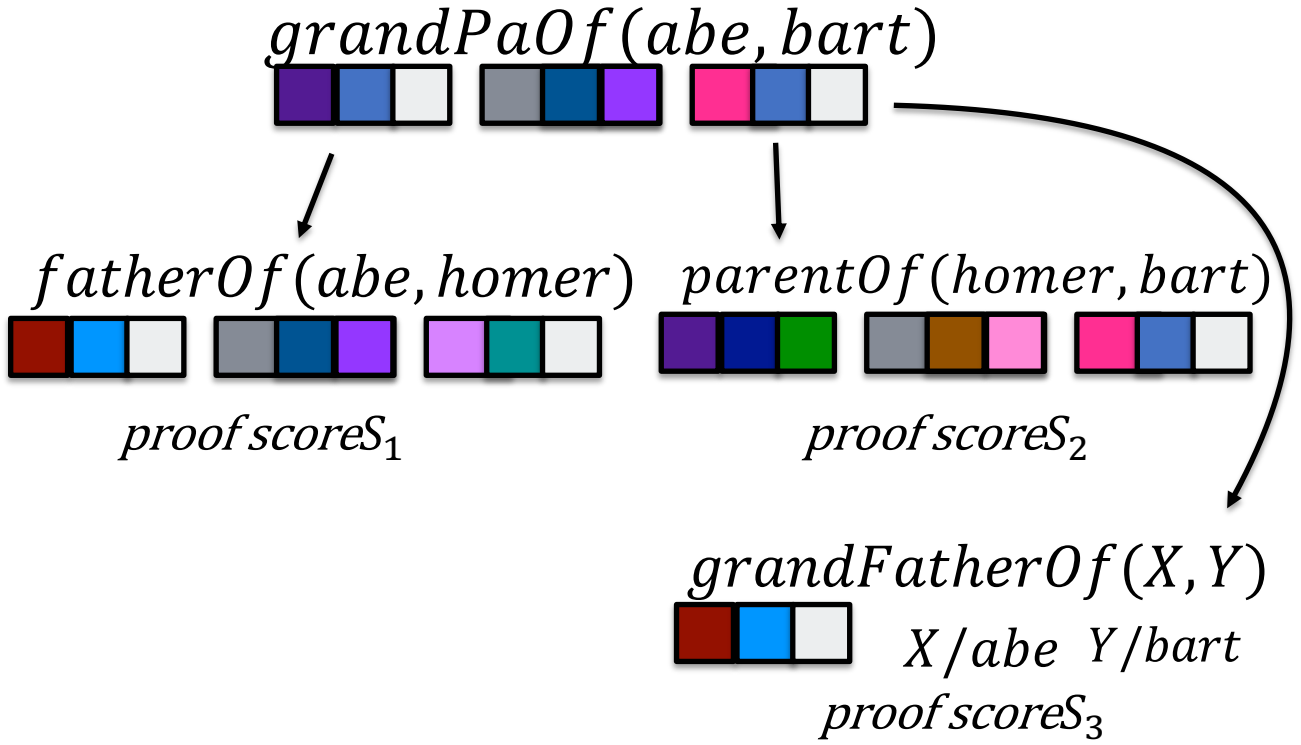


[Rocktäschel et al. 2017, Minervini et al. 2018, Welbl et al. 2019]

Backward Chaining — Differentiable Proving

Knowledge Base:

$fatherOf(abe, homer)$
$parentOf(homer, bart)$
$grandFatherOf(X, Y) \Leftarrow$
$fatherOf(X, Z),$
$parentOf(Z, Y).$



Subgoals:

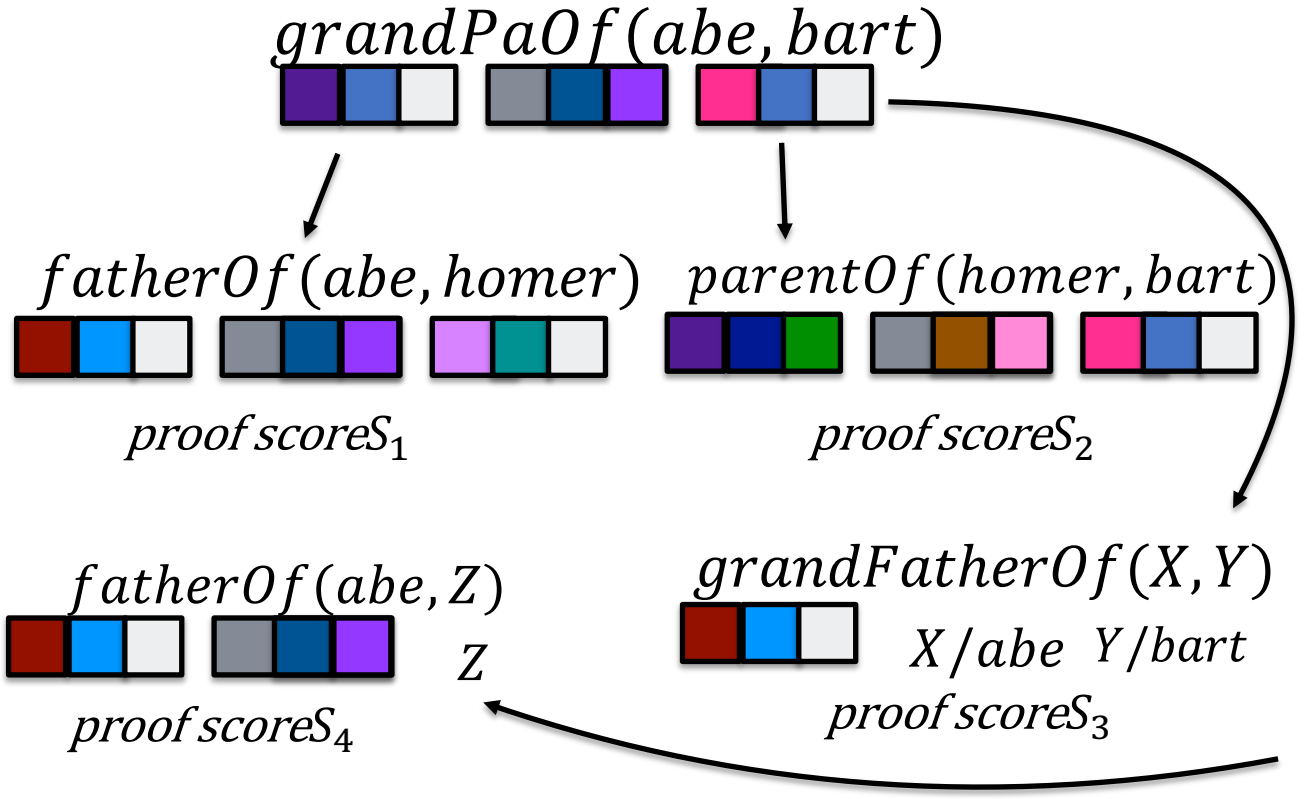
$fatherOf(abe, Z)$
$parentOf(Z, bart)$

[Rocktäschel et al. 2017, Minervini et al. 2018, Welbl et al. 2019]

Backward Chaining — Differentiable Proving

Knowledge Base:

$fatherOf(abe, homer)$
$parentOf(homer, bart)$
$grandFatherOf(X, Y) \Leftarrow$
$fatherOf(X, Z),$
$parentOf(Z, Y).$



Subgoals:

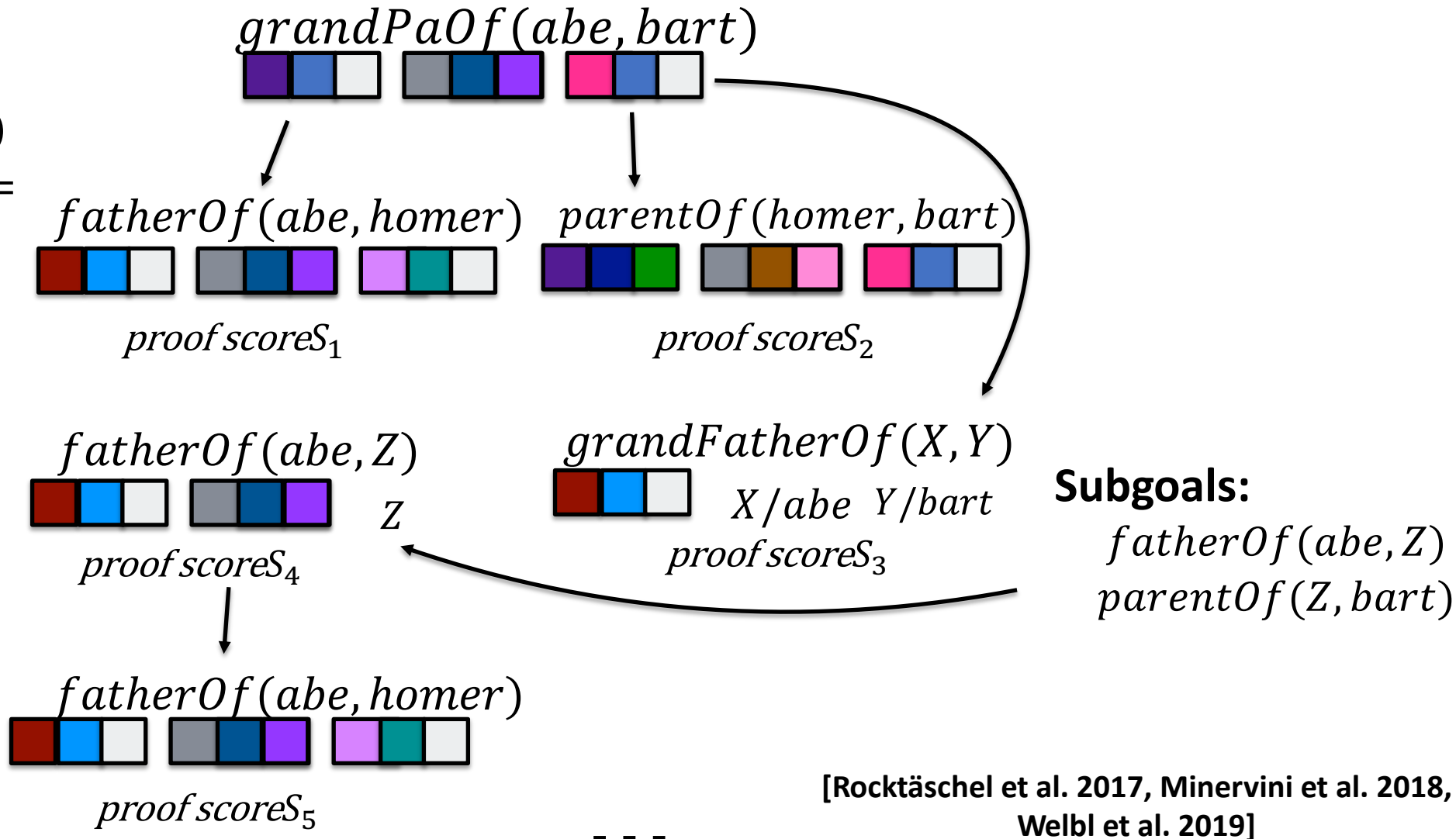
$fatherOf(abe, Z)$
$parentOf(Z, bart)$

[Rocktäschel et al. 2017, Minervini et al. 2018, Welbl et al. 2019]

Backward Chaining — Differentiable Proving

Knowledge Base:

$fatherOf(abe, homer)$
$parentOf(homer, bart)$
$grandFatherOf(X, Y) \Leftarrow$
$fatherOf(X, Z),$
$parentOf(Z, Y).$



Subgoals:

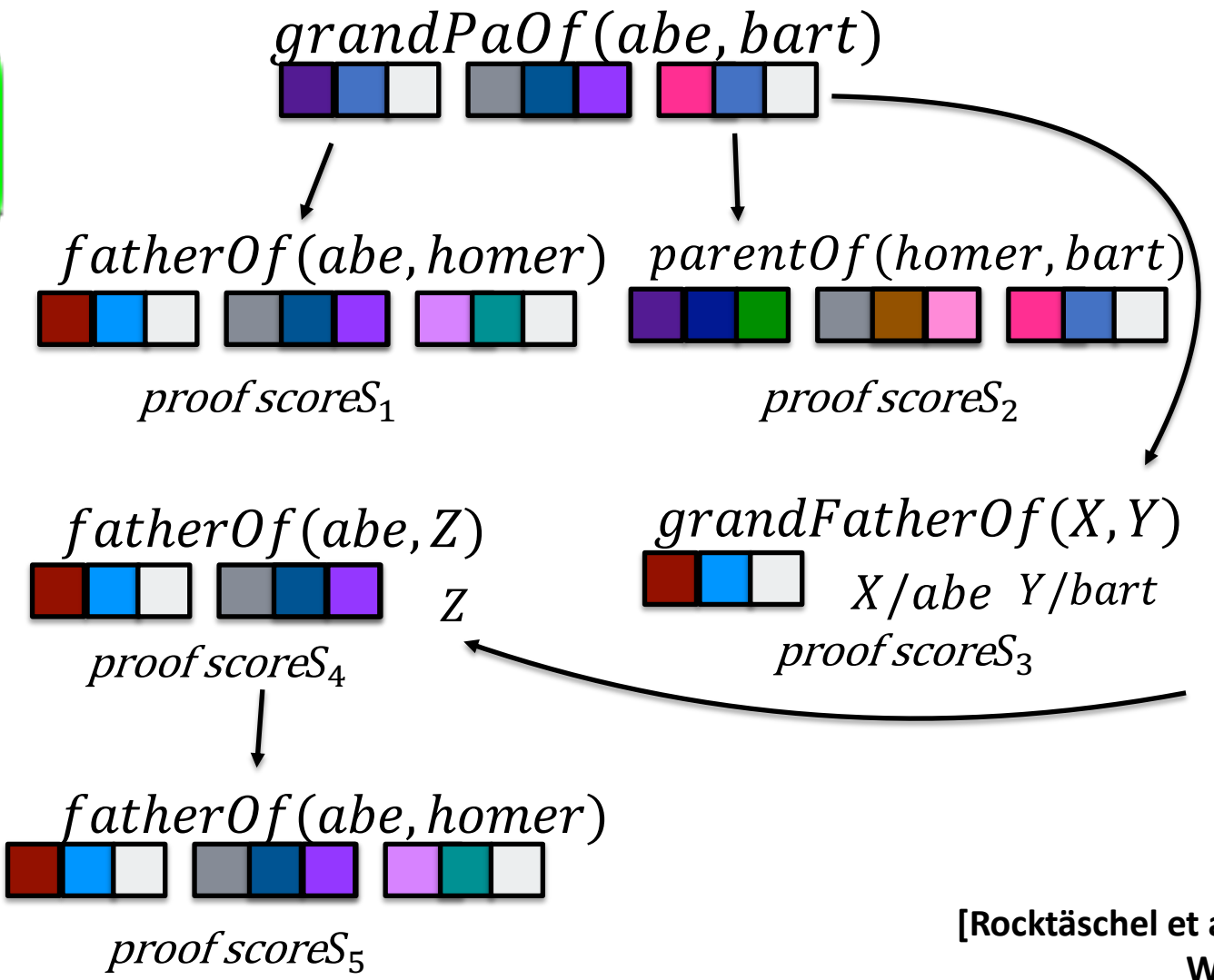
$fatherOf(abe, Z)$
$parentOf(Z, bart)$

[Rocktäschel et al. 2017, Minervini et al. 2018, Welbl et al. 2019]

Backward Chaining — Differentiable Proving

Knowledge Base:

fatherOf(abe, homer)
parentOf(homer, bart)
$\theta_1(X, Y) \Leftarrow \theta_2(X, Z), \theta_3(Z, Y)$



Subgoals:

$fatherOf(abe, Z)$
$parentOf(Z, bart)$

Training

Maximise Log-Likelihood:

$$\sum_{F \in K} \log p^{KB \setminus F}(F) - \sum_{\tilde{F} \sim corr(F)} \log p^{KB}(\tilde{F})$$

...

[Rocktäschel et al. 2017, Minervini et al. 2018, Welbl et al. 2019]

Differentiable Reasoning

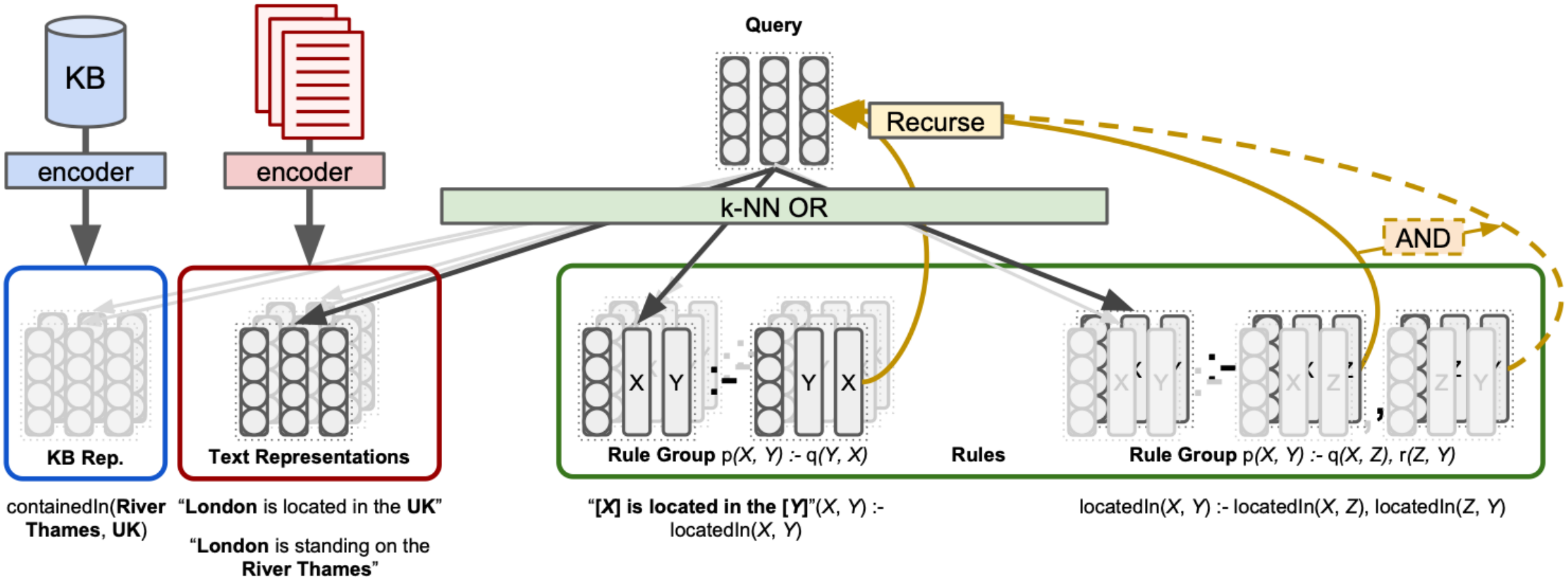
Corpus		Metric	Model			Examples of induced rules and their confidence
			ComplEx	**NTP**	**NTPλ**	
Countries	S1	AUC-PR	99.37 ± 0.4	90.83 ± 15.4	**100.00** $\pm$ 0.0	0.90 locatedIn(X,Y) :– locatedIn(X,Z), locatedIn(Z,Y).
	S2	AUC-PR	87.95 ± 2.8	87.40 ± 11.7	**93.04** $\pm$ 0.4	0.63 locatedIn(X,Y) :– neighborOf(X,Z), locatedIn(Z,Y).
	S3	AUC-PR	48.44 ± 6.3	56.68 ± 17.6	**77.26** $\pm$ 17.0	0.32 locatedIn(X,Y) :– neighborOf(X,Z), neighborOf(Z,W), locatedIn(W,Y).
Kinship		MRR	**0.81**	0.60	0.80	0.98 term15(X,Y) :– term5(Y,X)
		HITS@1	0.70	0.48	**0.76**	0.97 term18(X,Y) :– term18(Y,X)
		HITS@3	**0.89**	0.70	0.82	0.86 term4(X,Y) :– term4(Y,X)
		HITS@10	**0.98**	0.78	0.89	0.73 term12(X,Y) :– term10(X, Z), term12(Z, Y).
Nations		MRR	**0.75**	**0.75**	0.74	0.68 blockpositionindex(X,Y) :– blockpositionindex(Y,X).
		HITS@1	**0.62**	**0.62**	0.59	0.46 expeldiplomats(X,Y) :– negativebehavior(X,Y).
		HITS@3	0.84	0.86	**0.89**	0.38 negativecomm(X,Y) :– commonbloc0(X,Y).
		HITS@10	**0.99**	**0.99**	**0.99**	0.38 intergovorgs3(X,Y) :– intergovorgs(Y,X).
UMLS		MRR	0.89	0.88	**0.93**	0.88 interacts_with(X,Y) :– interacts_with(X,Z), interacts_with(Z,Y).
		HITS@1	0.82	0.82	**0.87**	0.77 isa(X,Y) :– isa(X,Z), isa(Z,Y).
		HITS@3	0.96	0.92	**0.98**	0.71 derivative_of(X,Y) :– derivative_of(X,Z), derivative_of(Z,Y).
		HITS@10	**1.00**	0.97	**1.00**	

Explainable Neural Link Prediction

	Query	Score S_ρ	Proofs / Explanations
WN18	part_of(CONGO.N.03, AFRICA.N.01)	0.995	part_of(X, Y) :- has_part(Y, X) has_part(AFRICA.N.01, CONGO.N.03)
		0.787	part_of(X, Y) :- instance_hyponym(Y, X) instance_hyponym(AFRICAN_COUNTRY.N.01, CONGO.N.03)
	hyponym(EXTINGUISH.V.04, DECOUPLE.V.03)	0.987	hyponym(X, Y) :- hypernym(Y, X) hypernym(DECOUPLE.V.03, EXTINGUISH.V.04)
		0.920	hypernym(SNUFF_OUT.V.01, EXTINGUISH.V.04)
	part_of(PITUITARY.N.01, DIENCEPHALON.N.01)	0.995	has_part(DIENCEPHALON.N.01, PITUITARY.N.01)
	has_part(TEXAS.N.01, ODESSA.N.02)	0.961	has_part(X, Y) :- part_of(Y, X) part_of(ODESSA.N.02, TEXAS.N.01)
	hyponym(SKELETAL_MUSCLE, ARTICULAR_MUSCLE)	0.987	hypernym(ARTICULAR_MUSCLE, SKELETAL_MUSCLE)
	deriv_related_form(REWRITE, REWRITING)	0.809	deriv_related_form(X, Y) :- hypernym(Y, X) hypernym(REVISE, REWRITE)
WN18RR	also_see(TRUE.A.01, FAITHFUL.A.01)	0.962	also_see(X, Y) :- also_see(Y, X) also_see(FAITHFUL.A.01, TRUE.A.01)
		0.590	also_see(CONSTANT.A.02, FAITHFUL.A.01)
	also_see(GOOD.A.03, VIRTUOUS.A.01)	0.962	also_see(VIRTUOUS.A.01, GOOD.A.03)
		0.702	also_see(RIGHTEOUS.A.01, VIRTUOUS.A.01)
	instance_hypernym(CHAPLIN, FILM_MAKER)	0.812	instance_hypernym(CHAPLIN, COMEDIAN)

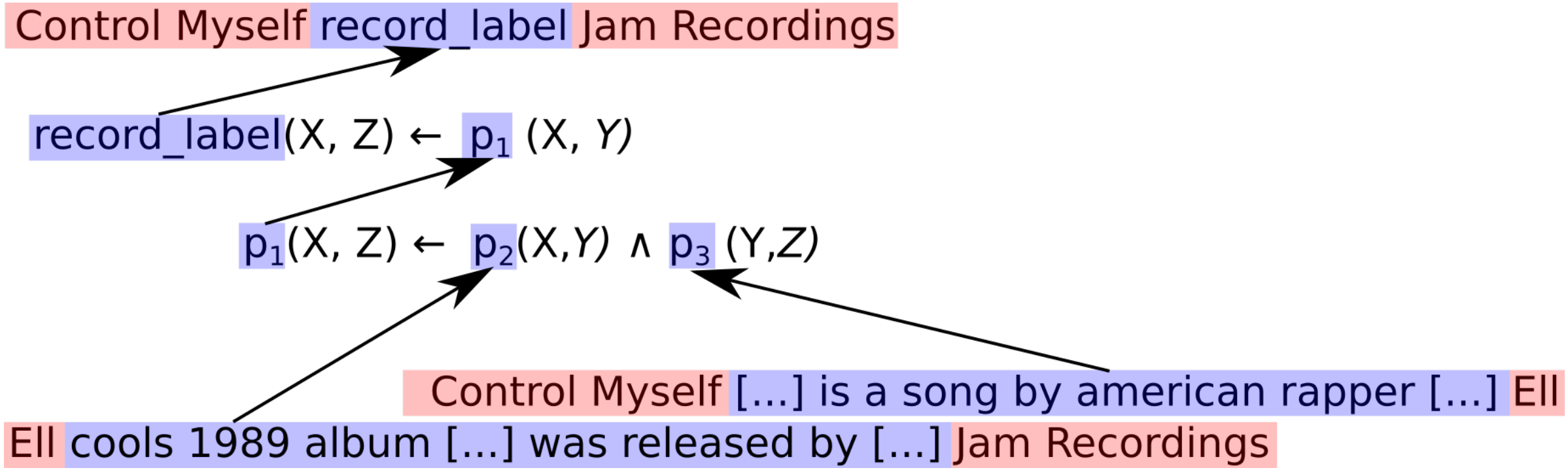
Reasoning Over Text

- We can embed facts from the KG and facts from text in a shared embedding space, and learn to reason over them jointly:



Reasoning Over Text

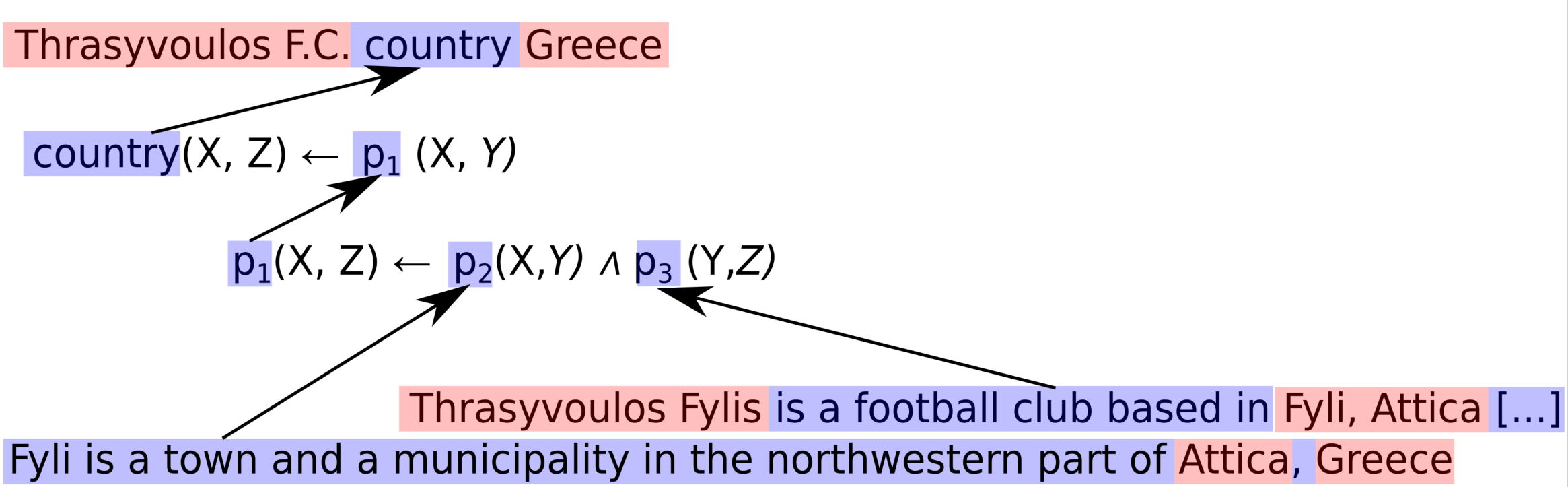
- We can embed facts from the KG and facts from text in a shared embedding space, and learn to reason over them jointly:



[Rocktäschel et al. 2017, Minervini et al. 2018, Welbl et al. 2019]

Reasoning Over Text

- We can embed facts from the KG and facts from text in a shared embedding space, and learn to reason over them jointly:



[Rocktäschel et al. 2017, Minervini et al. 2018, Welbl et al. 2019]

Neuro-Symbolic Integration — Recent Advances

- Recursive Reasoning Networks [Hohenecker et al. 2018] — given a OWL RL ontology, uses a differentiable model to update the entity and predicate representations.
- Deep ProbLog [Manhaeve et al. NeurIPS 2018] — extends the ProbLog probabilistic logic programming language with neural predicates that can be evaluated on e.g. sensory data (images, speech).
- Logic Tensor Networks [Serafini et al. 2016, 2017] — fully ground First Order Logic rules.
- AutoEncoder-like Architectures [Campero et al. 2018] — use end-to-end differentiable reasoning in the decoder of an autoencoder-like architecture to learn the minimal set of facts and rules that govern your domain via backprop.

References

- Maximilian Nickel, Kevin Murphy, Volker Tresp, Evgeniy Gabrilovich: **A Review of Relational Machine Learning for Knowledge Graphs**. Proceedings of the IEEE 104(1): 11-33 (2016)
- Lise Getoor and Ben Taskar: **Introduction to Statistical Relational Learning** (Adaptive Computation and Machine Learning). The MIT Press (2007)
- Xin Dong, Evgeniy Gabrilovich, Geremy Heitz, Wilko Horn, Ni Lao, Kevin Murphy, Thomas Strohmann, Shaohua Sun, Wei Zhang: **Knowledge vault: a web-scale approach to probabilistic knowledge fusion**. KDD 2014: 601-610
- Denis Krompaß, Stephan Baier, Volker Tresp: **Type-Constrained Representation Learning in Knowledge Graphs**. International Semantic Web Conference (1) 2015: 640-655
- L. A. Adamic and E. Adar: **Friends and neighbors on the Web**. Social Networks, vol. 25, no. 3, pp. 211–230, 2003
- A.-L. Barabási and R. Albert: **Emergence of Scaling in Random Networks**. Science, vol. 286, no. 5439, pp. 509–512, 1999
- L. Katz: **A new status index derived from sociometric analysis**. Psychometrika, vol. 18, no. 1, pp. 39–43, 1953
- E. A. Leicht, P. Holme, and M. E. Newman: **Vertex similarity in networks**. Physical Review E, vol. 73, no. 2, p. 026120, 2006
- S. Brin and L. Page: **The anatomy of a large-scale hypertextual Web search engine**. Computer networks and ISDN systems, vol. 30, no. 1, pp. 107–117, 1998.
- D. Liben-Nowell and J. Kleinberg: **The link-prediction problem for social networks**. Journal of the American society for information science and technology, vol. 58, no. 7, pp. 1019–1031, 2007.

References

- W. Liu and L. Lü: **Link prediction based on local random walk**. EPL (Europhysics Letters), vol. 89, no. 5, p. 58007, 2010.
- Stephen Muggleton: **Inverting Entailment and Progol**. Machine Intelligence 14 1993: 135-190
- Ashwin Srinivasan: **The Aleph Manual**. http://www.di.ubi.pt/~jpaulo/competence/tutorials/aleph.pdf 1999
- Jens Lehmann: **DL-Learner: Learning Concepts in Description Logics**. Journal of Machine Learning Research 10: 2639-2642 (2009)
- J. R. Quinlan: **Learning logical definitions from relations**. Machine Learning, vol. 5, pp. 239–266, 1990
- Ni Lao, Tom M. Mitchell, William W. Cohen: **Random Walk Inference and Learning in A Large Scale Knowledge Base**. EMNLP 2011: 529-539
- Luis Galárraga, Christina Teflioudi, Katja Hose, Fabian M. Suchanek: **Fast rule mining in ontological knowledge bases with AMIE+.** VLDB J. 24(6): 707-730 (2015)
- Maximilian Nickel, Volker Tresp, Hans-Peter Kriegel: **A Three-Way Model for Collective Learning on Multi-Relational Data**. ICML 2011: 809-816
- Antoine Bordes, Nicolas Usunier, Alberto García-Durán, Jason Weston, Oksana Yakhnenko: **Translating Embeddings for Modeling Multi-relational Data**. NIPS 2013: 2787-2795
- Bishan Yang, Wen-tau Yih, Xiaodong He, Jianfeng Gao, Li Deng: **Embedding Entities and Relations for Learning and Inference in Knowledge Bases**. CoRR abs/1412.6575 (2014)

References

- Maximilian Nickel, Lorenzo Rosasco, Tomaso A. Poggio: **Holographic Embeddings of Knowledge Graphs**. AAAI 2016: 1955-1961
- Théo Trouillon, Johannes Welbl, Sebastian Riedel, Éric Gaussier, Guillaume Bouchard: **Complex Embeddings for Simple Link Prediction**. ICML 2016: 2071-2080
- Tim Dettmers, Pasquale Minervini, Pontus Stenetorp, Sebastian Riedel: **Convolutional 2D Knowledge Graph Embeddings**. AAAI 2018: 1811-1818
- Timothée Lacroix, Nicolas Usunier, Guillaume Obozinski: **Canonical Tensor Decomposition for Knowledge Base Completion**. ICML 2018: 2869-2878
- Pasquale Minervini, Luca Costabello, Emir Muñoz, Vít Novácek, Pierre-Yves Vandenbussche: **Regularizing Knowledge Graph Embeddings via Equivalence and Inversion Axioms**. ECML/PKDD (1) 2017: 668-683
- Pasquale Minervini, Thomas Demeester, Tim Rocktäschel, Sebastian Riedel: **Adversarial Sets for Regularising Neural Link Predictors**. UAI 2017
- Maximilian Nickel, Xueyan Jiang, Volker Tresp: **Reducing the Rank in Relational Factorization Models by Including Observable Patterns**. NIPS 2014: 1179-1187
- Richard Evans, Edward Grefenstette: **Learning Explanatory Rules from Noisy Data**. J. Artif. Intell. Res. 61: 1-64 (2018)
- Tim Rocktäschel, Sebastian Riedel: **End-to-end Differentiable Proving**. NeurIPS 2017: 3791-3803
- Patrick Hohenecker, Thomas Lukasiewicz: **Ontology Reasoning with Deep Neural Networks**. CoRR abs/1808.07980 (2018)

References

- Pasquale Minervini, Matko Bosnjak, Tim Rocktäschel, Sebastian Riedel: **Towards Neural Theorem Proving at Scale**. CoRR abs/1807.08204 (2018)
- Leon Weber, Pasquale Minervini, Jannes Münchmeyer, Ulf Leser, Tim Rocktäschel: **NLProlog: Reasoning with Weak Unification for Question Answering in Natural Language**. ACL (1)2019: 6151-6161
- Robin Manhaeve, Sebastijan Dumancic, Angelika Kimmig, Thomas Demeester, Luc De Raedt: **DeepProbLog: Neural Probabilistic Logic Programming**. NeurIPS 2018: 3753-3763
- Luciano Serafini, Artur S. d'Avila Garcez: **Logic Tensor Networks: Deep Learning and Logical Reasoning from Data and Knowledge**. CoRR abs/1606.04422 (2016)
- Ivan Donadello, Luciano Serafini, Artur S. d'Avila Garcez: **Logic Tensor Networks for Semantic Image Interpretation**. IJCAI 2017: 1596-1602
- Andres Campero, Aldo Pareja, Tim Klinger, Josh Tenenbaum, Sebastian Riedel: **Logical Rule Induction and Theory Learning Using Neural Theorem Proving**. CoRRabs/1809.02193
- Georgina Peake, Jun Wang: **Explanation Mining: Post Hoc Interpretability of Latent Factor Models for Recommendation Systems**. KDD 2018: 2060-2069
- Arthur Colombini Gusmão, Alvaro Henrique Chaim Correia, Glauber De Bona, Fábio Gagliardi Cozman: **Interpreting Embedding Models of Knowledge Bases: A Pedagogical Approach**. CoRR abs/1806.09504 (2018)

- Iván Sánchez Carmona, Sebastian Riedel: **Extracting Interpretable Models from Matrix Factorization Models**. CoCo@NIPS 2015
- Vicente Iván Sánchez Carmona, Tim Rocktäschel, Sebastian Riedel, Sameer Singh: **Towards Extracting Faithful and Descriptive Representations of Latent Variable Models**. AAAI Spring Symposia 2015
- Tareq B. Malas et al.: **Drug prioritization using the semantic properties of a knowledge graph**. Nature 2019
- Freddy Lecue: **On The Role of Knowledge Graphs in Explainable AI**. Semantic Web Journal 2019
- Cataldo Musto, Fedelucio Narducci, Pasquale Lops, Marco de Gemmis, Giovanni Semeraro: **Linked open data-based explanations for transparent recommender systems**. Int. J. Hum.-Comput. Stud. 121: 93-107 (2019)
- Tommaso Di Noia, Roberto Mirizzi, Vito Claudio Ostuni, Davide Romito, Markus Zanker**: Linked open data to support content-based recommender systems**. I-SEMANTICS 2012: 1-8
- Vito Claudio Ostuni, Tommaso Di Noia, Eugenio Di Sciascio, Roberto Mirizzi: **Top-N recommendations from implicit feedback leveraging linked open dat**a. RecSys 2013: 85-92
- David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, Antonio Torralba**: Network Dissection: Quantifying Interpretability of Deep Visual Representations**. CVPR 2017: 3319-3327

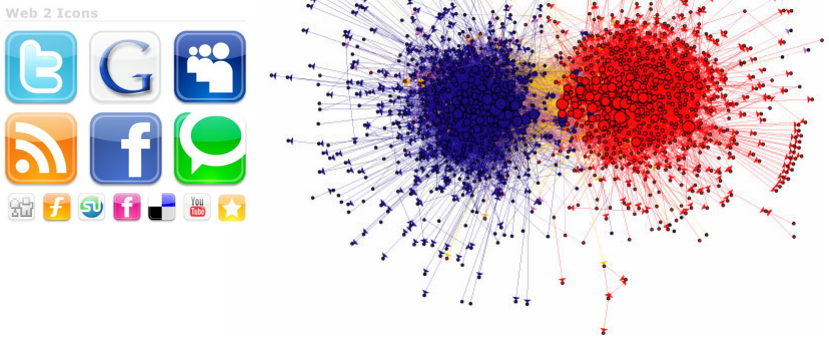
Explaining Privacy Risks

Anna Monreale, University of Pisa, Pisa

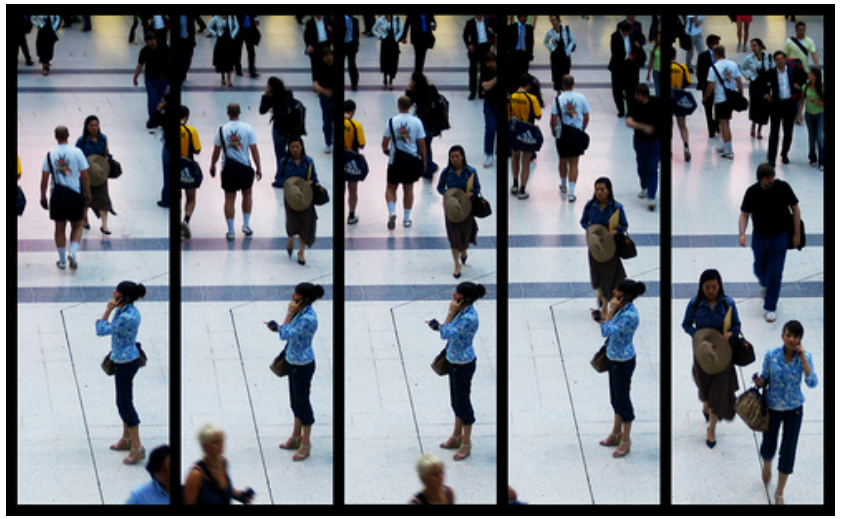
Big data “proxies” of social life

Shopping patterns & lifestyle

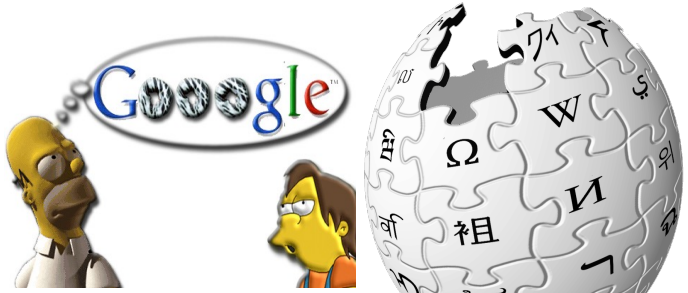
Relationships & social ties



Movements

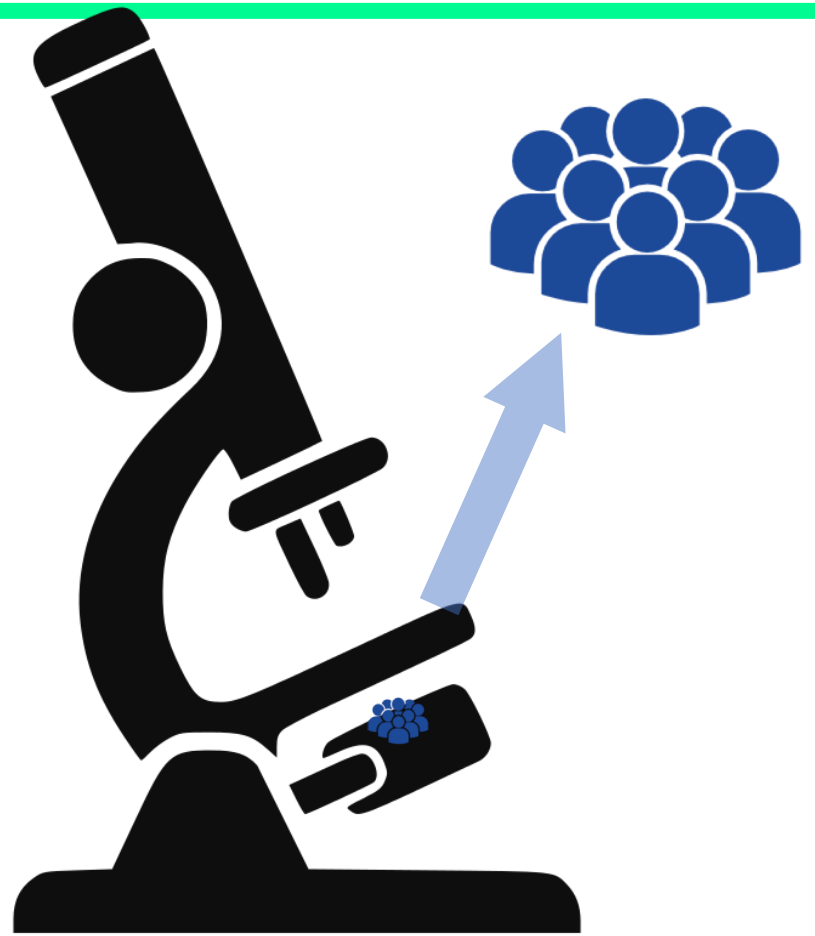


Desires, opinions, sentiments



WIKIPEDIA
The Free Encyclopedia

Big Data Analytics & Social Mining



The main tool for a Data Scientist to measure, understand, and possibly predict human behavior

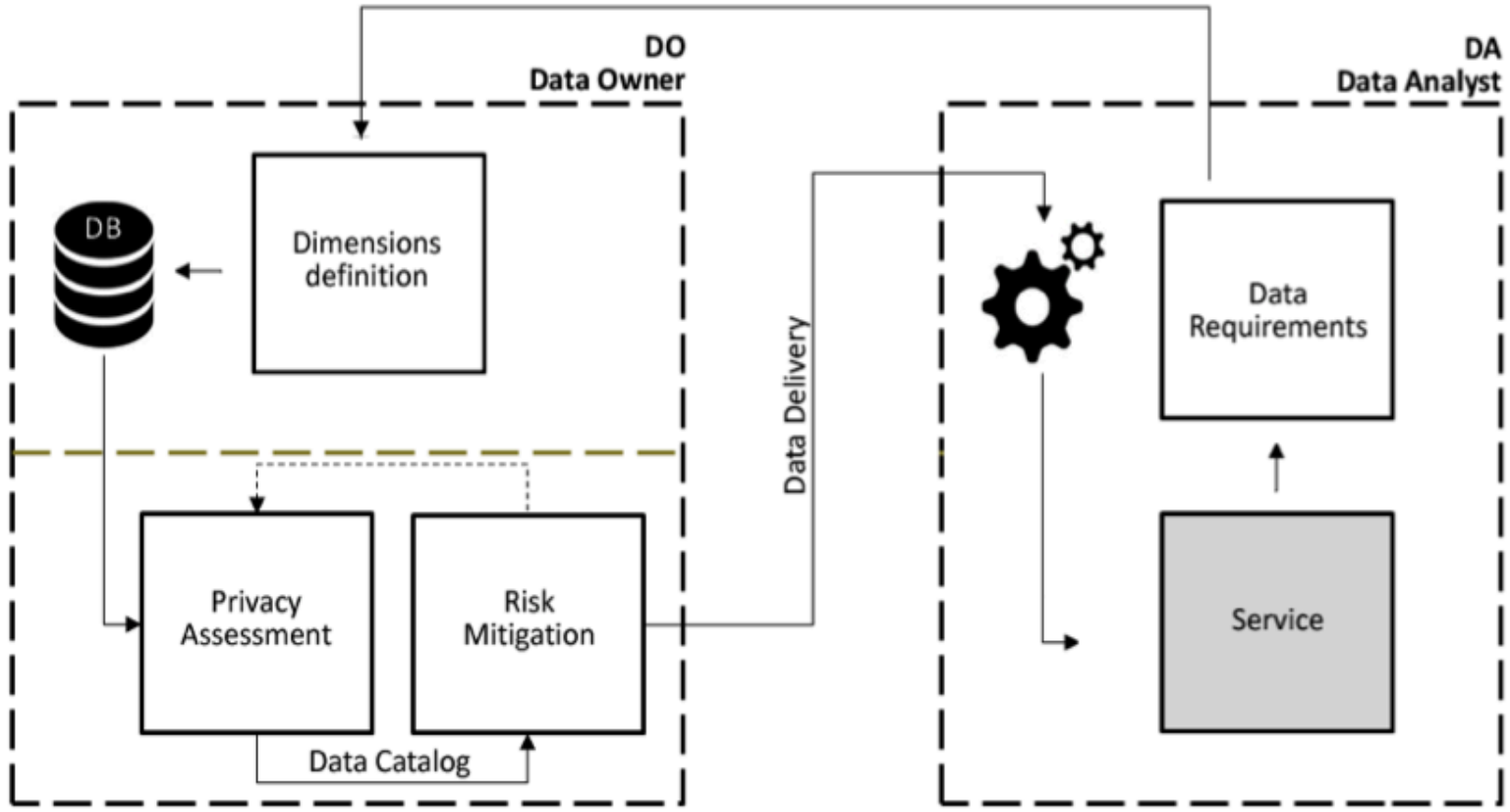
Data Scientist needs to take into account ethical and legal aspects and social impact of data science

Data Protection & Privacy

- **A fundamental human right**
- Any individual has the **right to a private life**, to be autonomous, to **control** information about yourself
- Any individual has the right to privacy protection
  - The right to be **directly or indirectly non-identifiable**
- Any data processing (es: data mining, data analysis, AI, ML, …) on this kind of data can bring to **individual privacy violation**

Privacy by Design & Risk Assessment

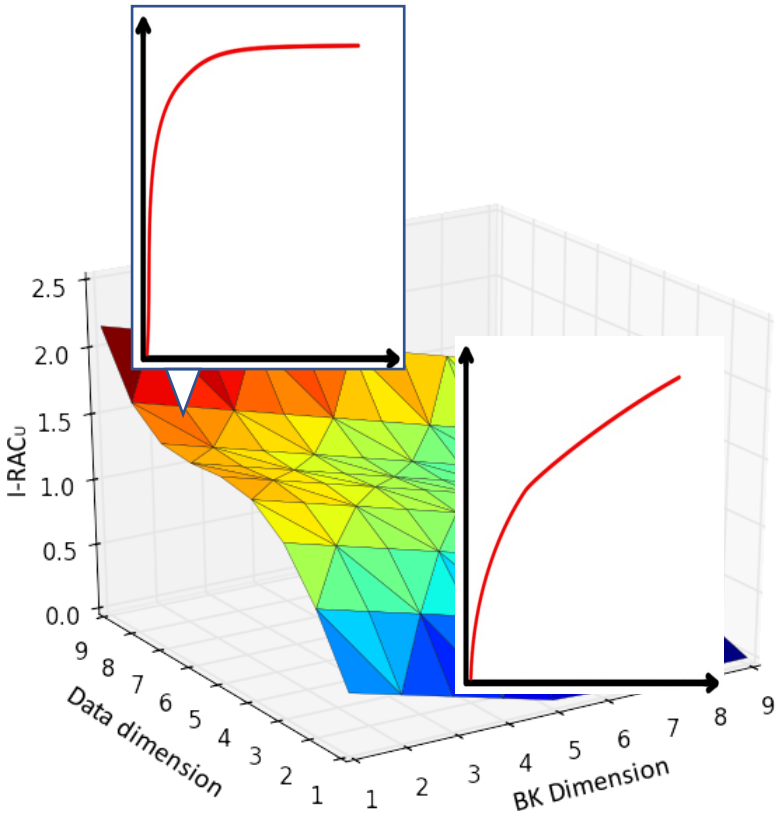
Privacy Risk Assessment Framework for Data Sharing



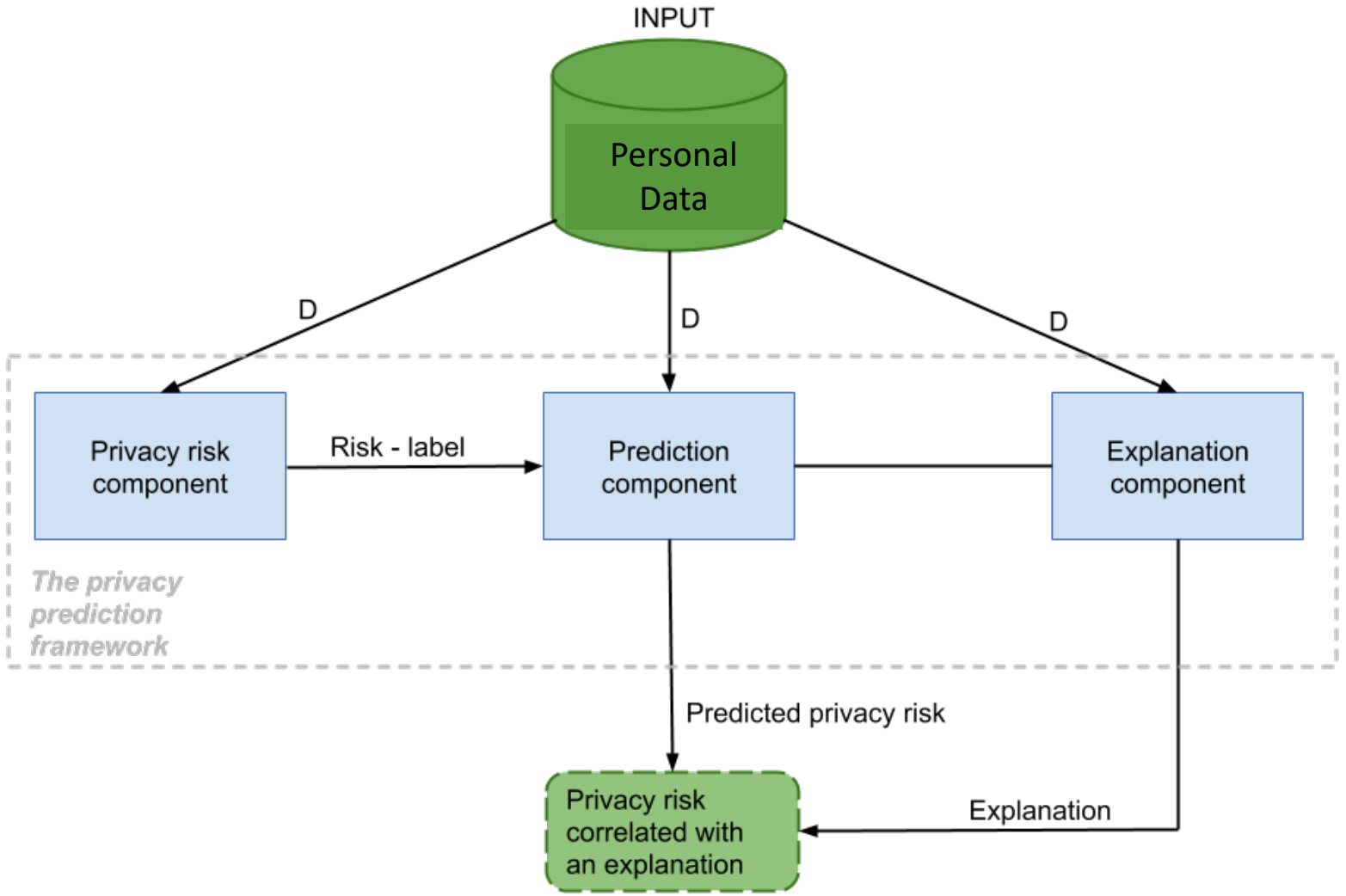
F. Pratesi, A. Monreale, R. Trasarti, F. Giannotti, D. Pedreschi, T. Yanagihara: **PRUDEnce: a System for Assessing Privacy Risk vs Utility in Data Sharing Ecosystems**. Transactions on Data Privacy 11(2): 139-167 (2018)

Privacy Risk Assessment Framework for Data Sharing

- Data Catalog
  - For each:
    - Data Format, i.e., the data needed for the service
    - Risk Assessment Setting, i.e., the set of pre-processing and privacy attacks
  - The Data Catalog provides:
    - **Quantification of Privacy Risk**, i.e., the evaluation of the real risk of re-identification
    - **Quantification of Data Quality**, i.e., the quality level we can achieve with private data, compared with the data quality of original data.

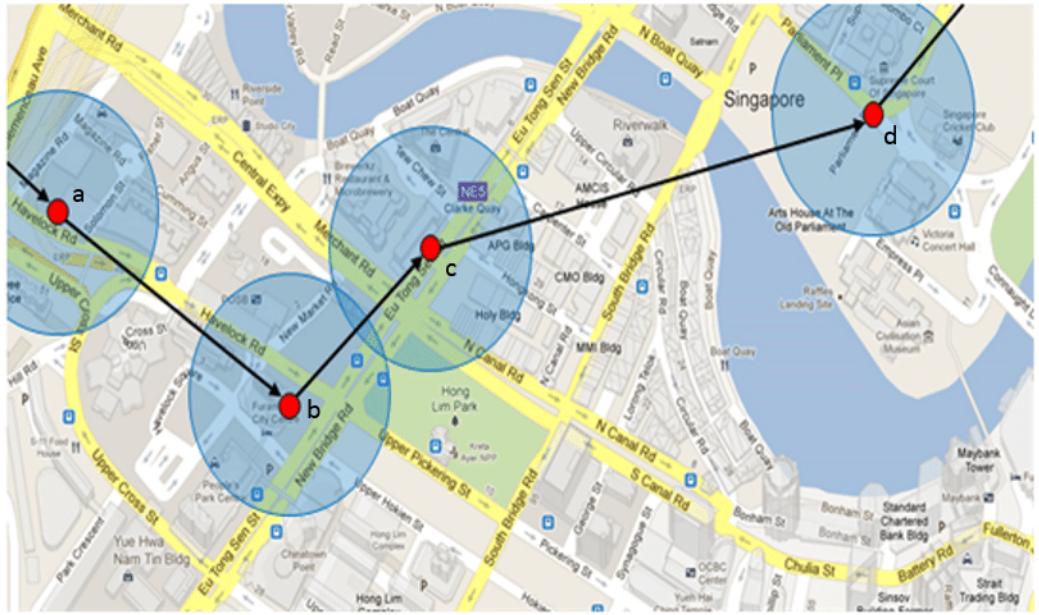


Privacy Risk Prediction



Sequence Data

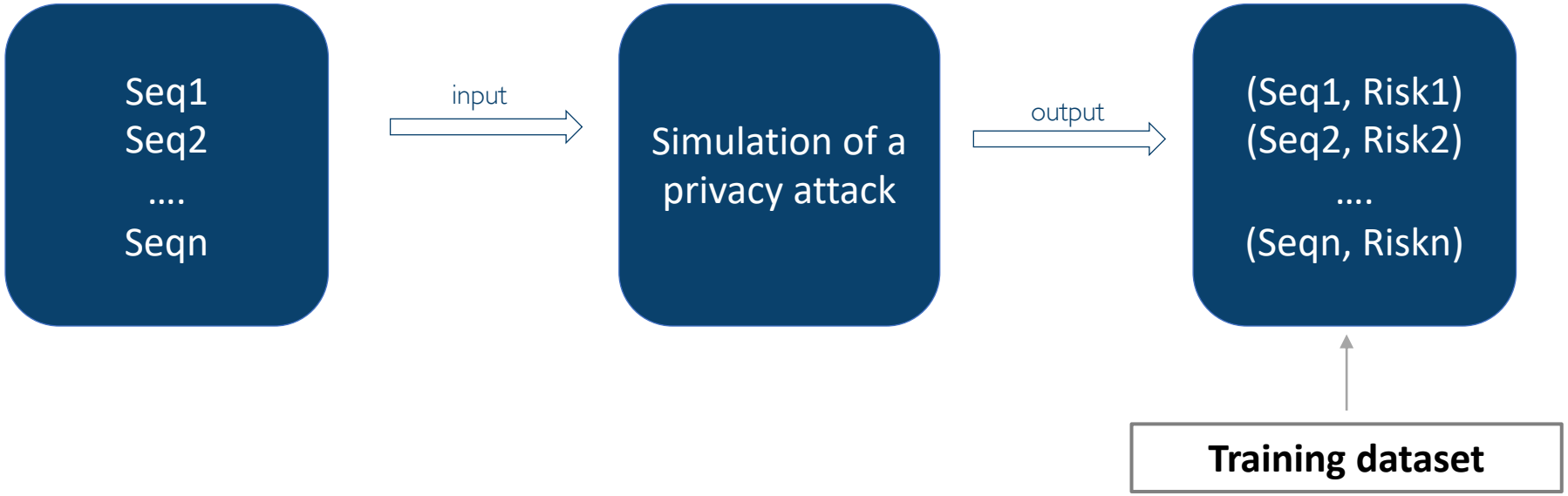
Mobility data



Retail data

Privacy Risk Component

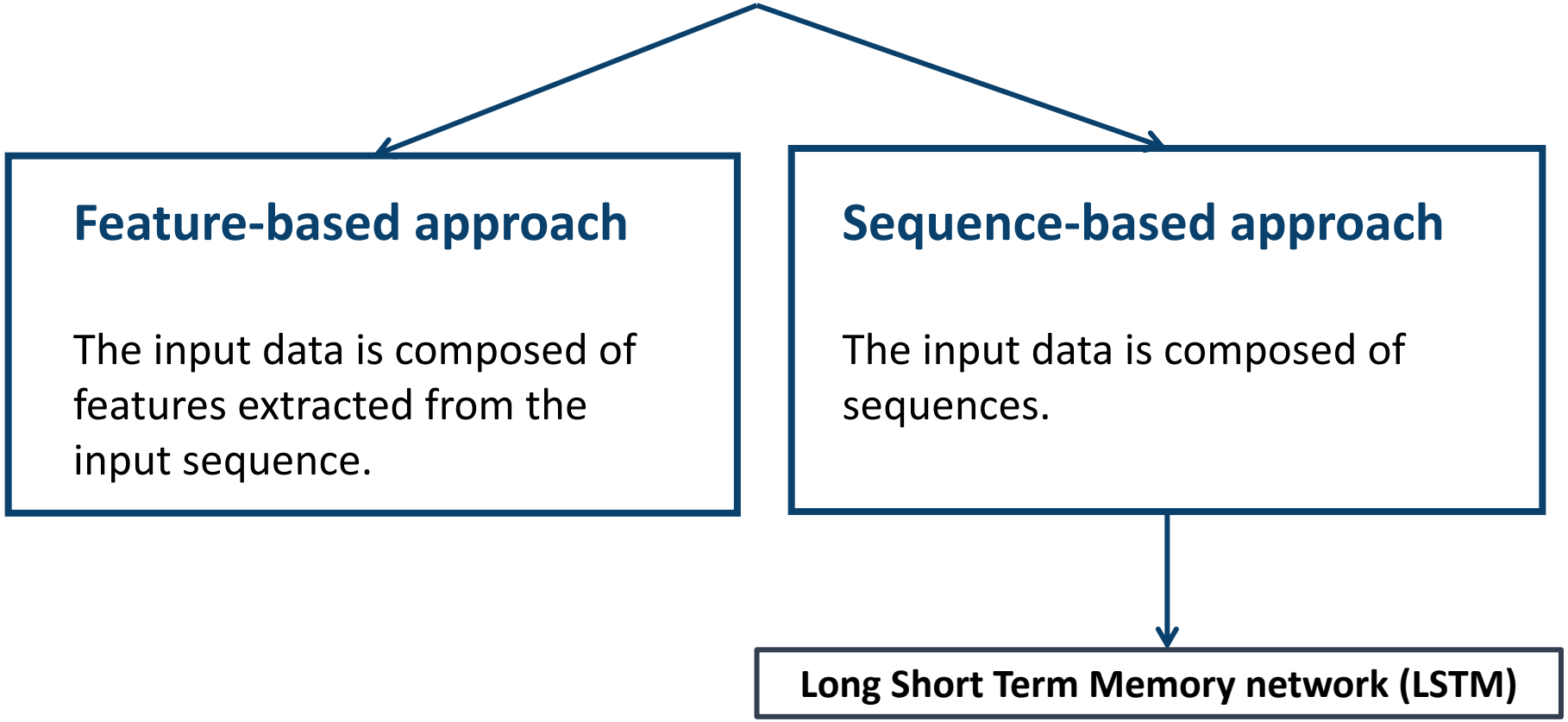
TASK: it provides the target output for the machine learning models.



Predictor Component

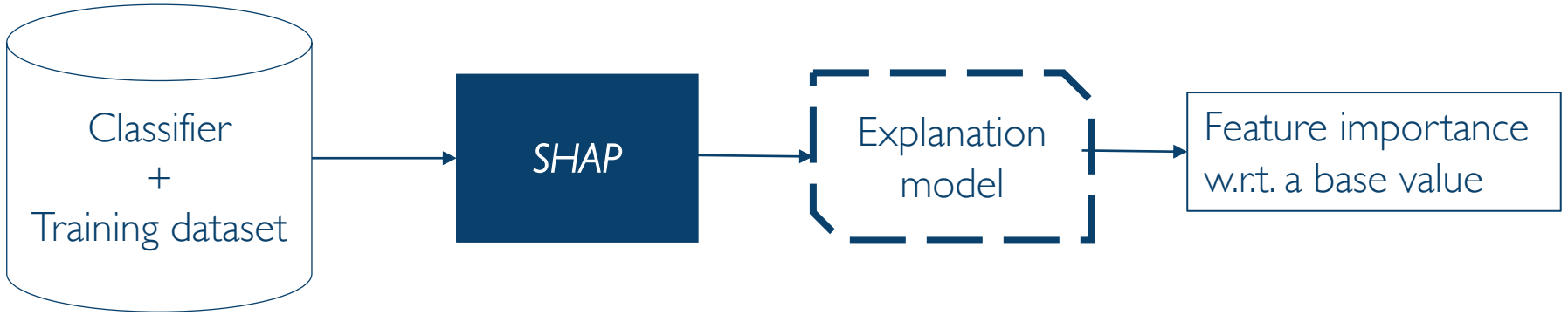
TASK: it predicts the privacy risk for each sequence.

Two approaches:



Explanation Component

TASK: it provides an explanation about the reasoning of the machine learning model.



SHAP: Shapley Additive Explanation

- **Game Theory:** Branch of micro-economics dealing with interactions between decision-making agents.
- **Cooperative Game Theory:** Sub-field of game theory where players are “working together” to achieve a common goal.
- In Machine Learning:
  - **game** is the prediction task for a single instance
  - **gain** is the **actual prediction** for this instance minus the **average prediction of all instances**
  - **players** are the **feature values** of the instance, which collaborate to receive the gain

Shapley Value

- **Key Idea:** Measure each player's contribution to the team's outcome.

- **Heuristic:** If we remove a player from the team and the outcome doesn't change, then the **player wasn't useful.**

Shapley Value

Intuition

For each player compute each outcome where the player was present and compare it to the outcome where the player was not present

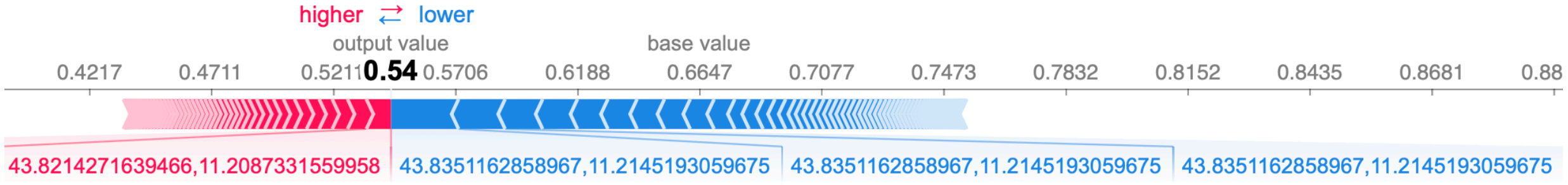
For each feature i:

- **Average of all possible differences between predictions of the model without feature i, and the ones with feature i**
- Computation of each coalition with feature i

SHAP Explanation

Retail data

Mobility data

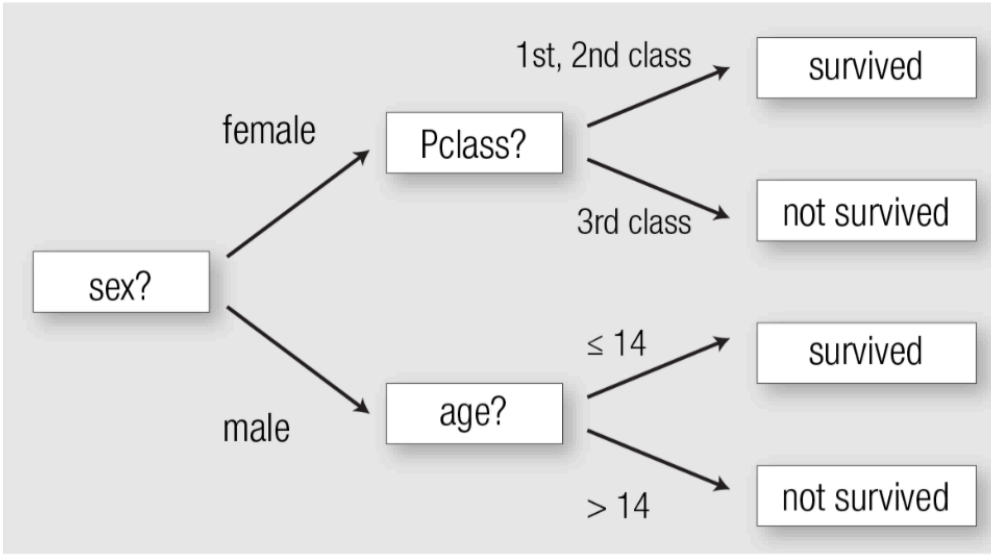


Visualizing Explanations

Riccardo Guidotti, Salvo Rinzivillo, ISTI-CNR, Pisa

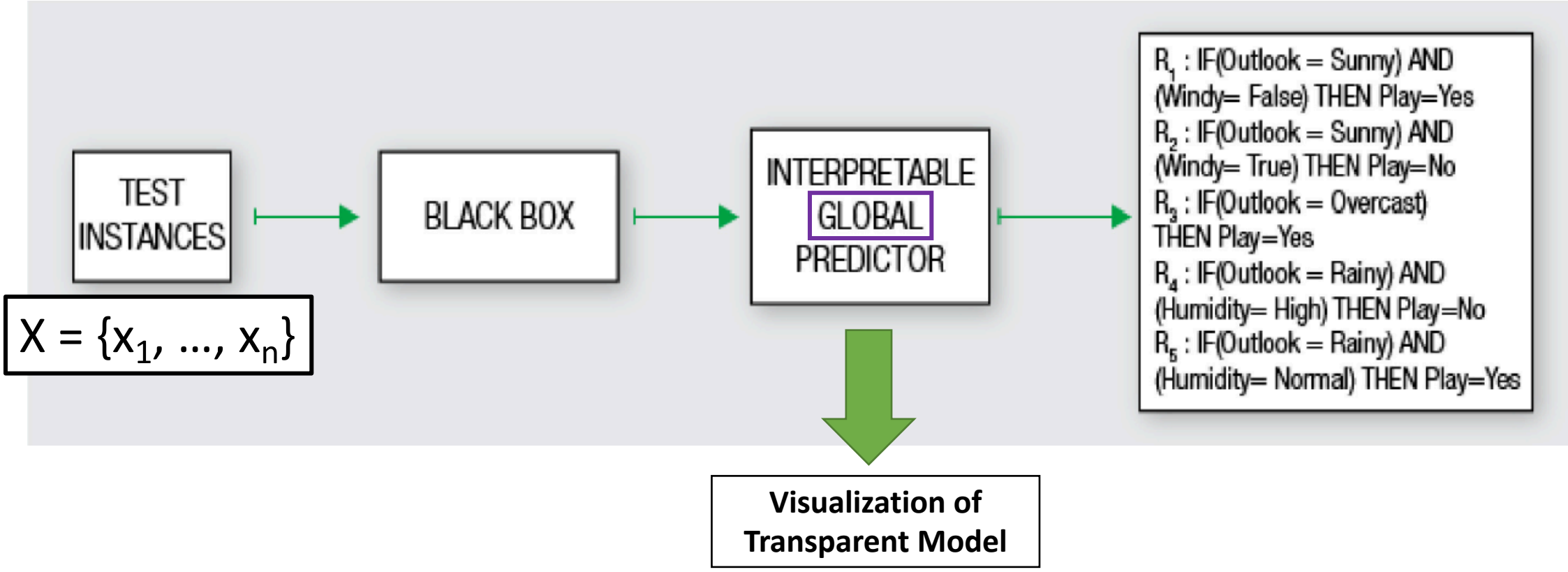
Transparent Model Visualization

- Representation of model on visual space
- Pro
  - Vision at a glance
  - Reduction of complexity through interaction
  - Exploits visual clues to highlight relevant patterns and properties
- Cons
  - Complex visualization when model is too large
  - Low efficiency of visual space occupancy (e.g. decision trees)

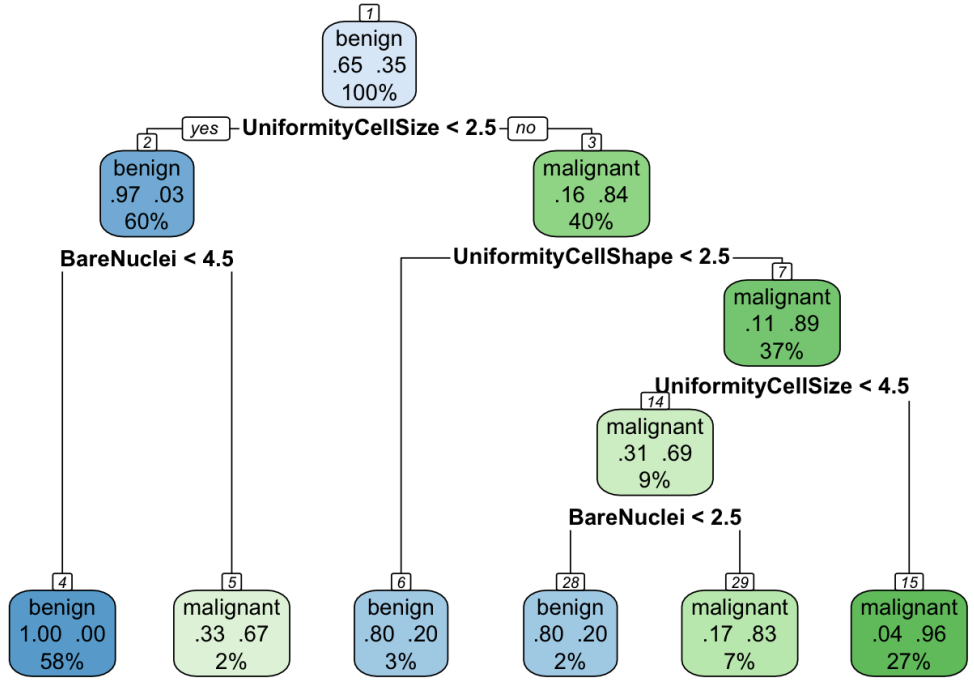


Model Explanation Problem to Visualization

In case of BB models, derive an interpretable/transparent box to be visualized.



Decision Tree and Rules Visualization

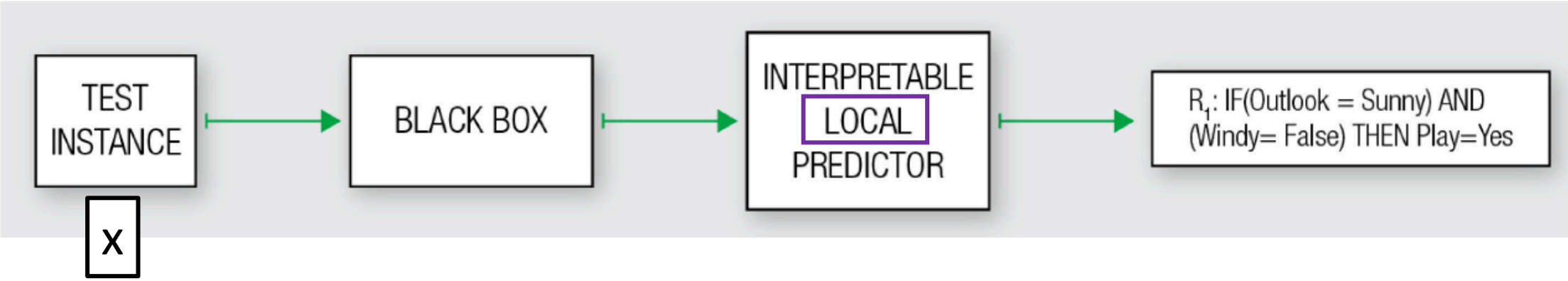


Mark Craven and JudeW. Shavlik. 1996. ***Extracting tree-structured representations of trained networks***. NIPS.

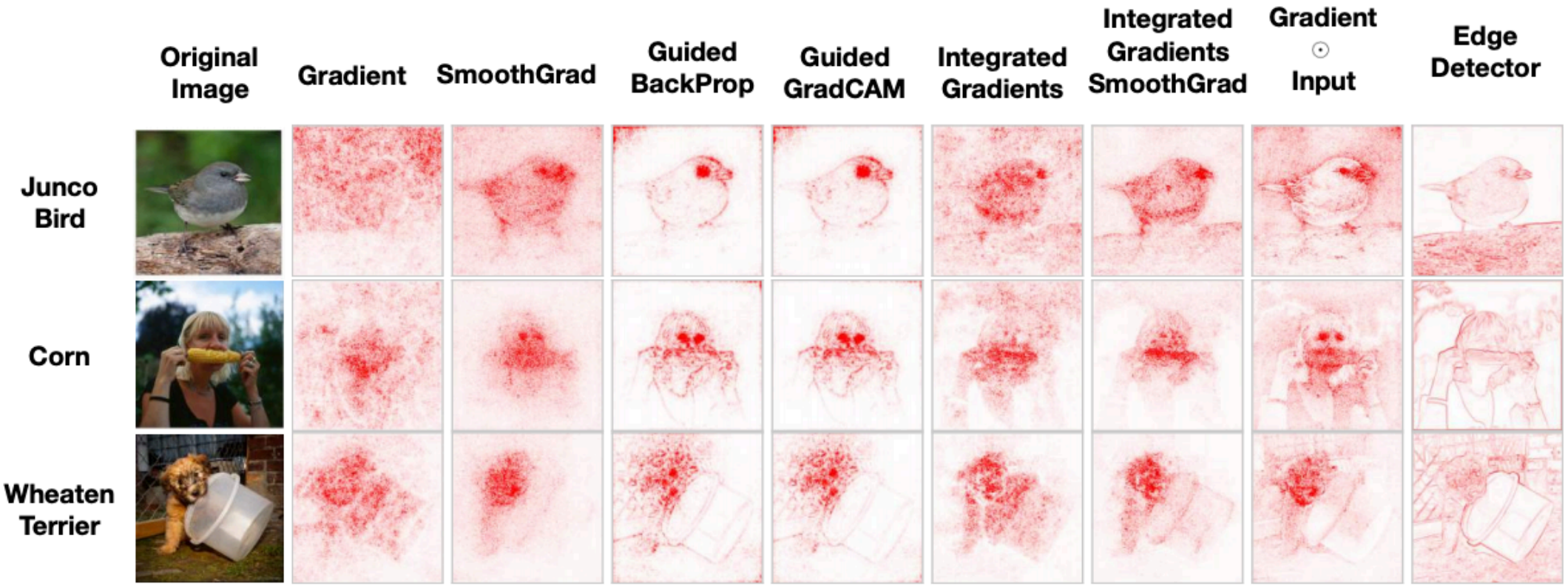
Yao Ming, Huamin Qu, and Enrico Bertini. **RuleMatrix: Visualizing and Understanding Classifiers with Rules**. IEEE Transactions on Visualization and Computer Graphics, 2019

Outcome Explanation Problem to Visualization

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



Feature Relevance on the Input Space



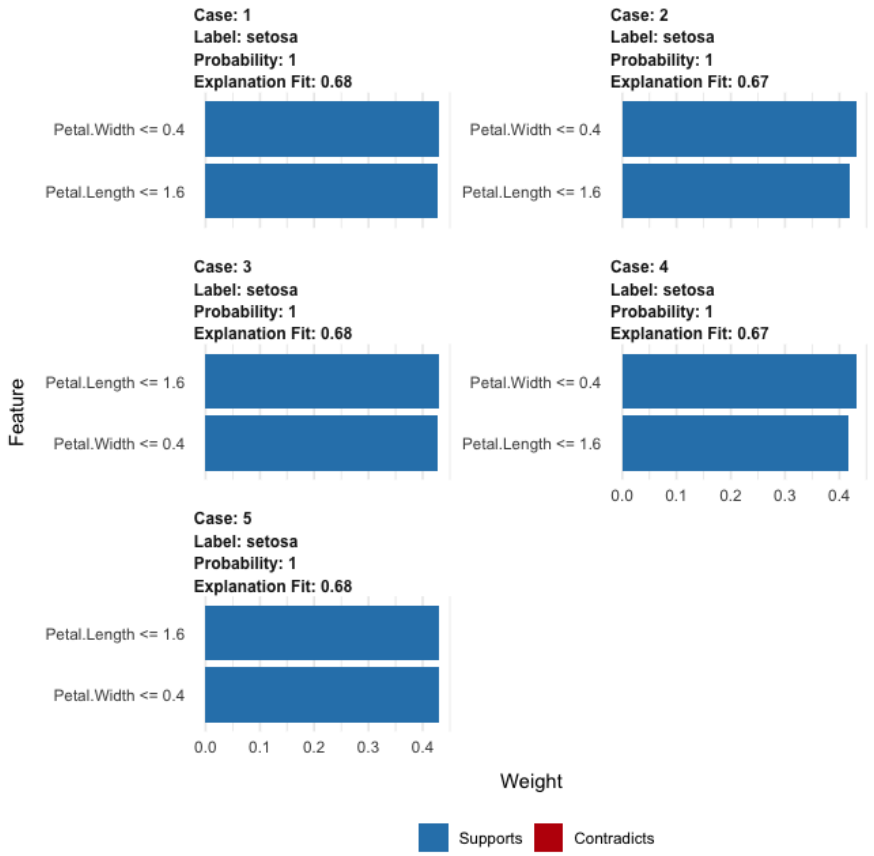
Julius Adebayo, Justin Gilmer, Michael Christoph Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. 2018.

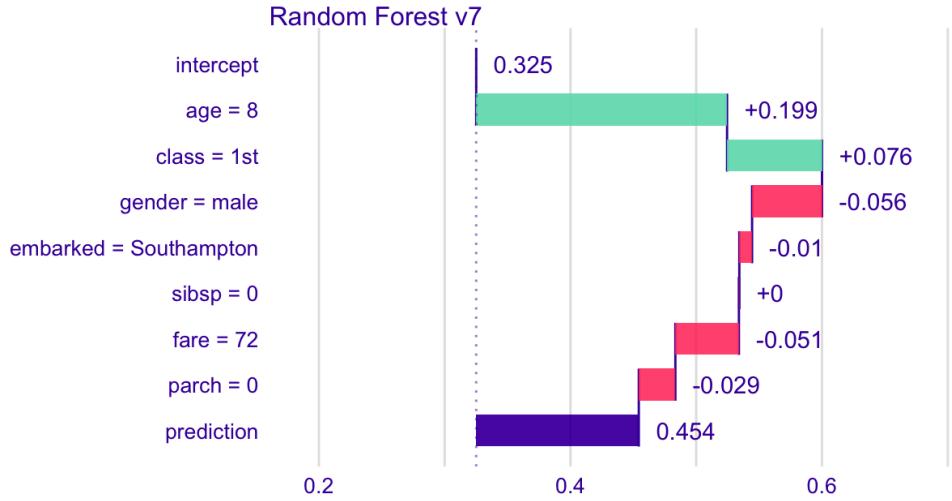
Feature Relevance on the Input Space

User 156	*Sentence level attractiveness*	**Election is a 1999 American comedy-drama film directed and written by Alexander Payne and adapted by him and Jim Taylor from Tom Perrotta's 1998 novel of the same title.** The plot revolves around a high school election and satirizes both suburban high school life and politics. **The film stars Matthew Broderick as Jim McAllister, a popular high school social studies teacher in suburban Omaha, Nebraska, and Reese Witherspoon as Tracy Flick, around the time of the school's student body election. When Tracy qualifies to run for class president, McAllister believes she does not deserve the title and tries his best to stop her from winning. Election opened to acclaim from critics, who praised its writing and direction.** The film received an Academy Award nomination for Best Adapted Screenplay, a Golden Globe nomination for Witherspoon in the Best Actress category, and the Independent Spirit Award for Best Film in 1999.
	Word level attractiveness	Election is a 1999 American **comedy-drama** film directed and written by Alexander Payne and adapted by him and Jim Taylor from Tom Perrotta's 1998 novel of the same title.
	Cast member attractiveness	**Alexander Payne**, Reese Witherspoon, Matthew Broderick, Jim Taylor
User 2163	*Sentence level attractiveness*	Election is a 1999 American comedy-drama film directed and written by Alexander Payne and adapted by him and Jim Taylor from Tom Perrotta's 1998 novel of the same title. The plot revolves around a high school election and satirizes both suburban high school life and politics. The film stars Matthew Broderick as Jim McAllister, a popular high school social studies teacher in suburban Omaha, Nebraska, and Reese Witherspoon as Tracy Flick, around the time of the school's student body election. When Tracy qualifies to run for class president, McAllister believes she does not deserve the title and tries his best to stop her from winning. Election opened to acclaim from critics, who praised its writing and direction. **The film received an Academy Award nomination for Best Adapted Screenplay, a Golden Globe nomination for Witherspoon in the Best Actress category, and the Independent Spirit Award for Best Film in 1999.**
	Word level attractiveness	The film received an Academy **Award** nomination for **Best** Adapted Screenplay, a Golden Globe nomination for Witherspoon in the **Best** Actress category, and the Independent Spirit **Award** for **Best** Film in 1999
	Cast member attractiveness	Alexander Payne, **Reese Witherspoon**, Matthew Broderick, Jim Taylor

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

Feature Properties and Relevance



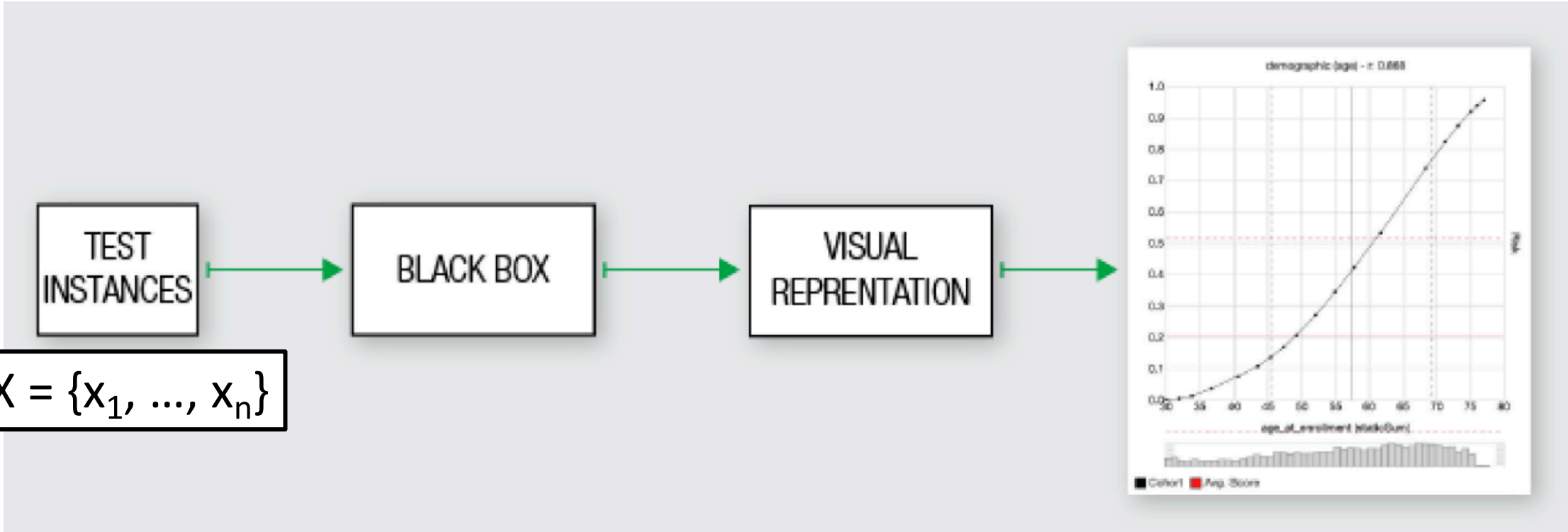


Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. **Why should i trust you?: Explaining the predictions of any classifier**. KDD.

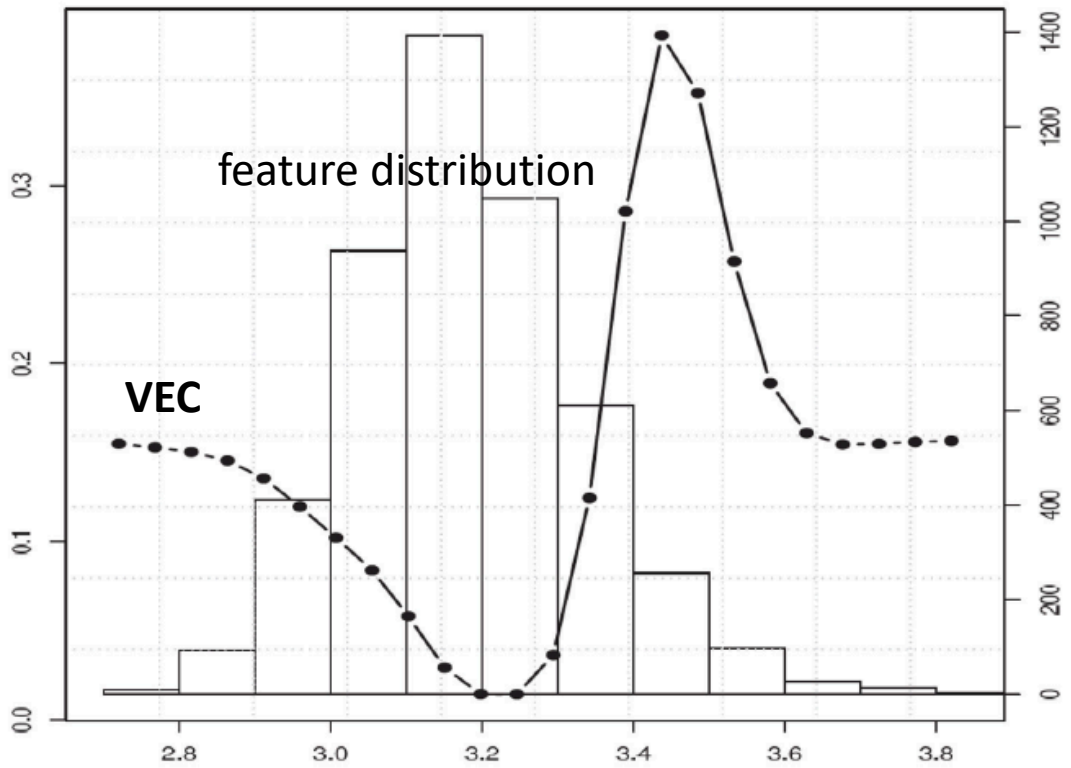
Gosiewska A, Biecek P (2019). "iBreakDown: Uncertainty of Model Explanations for Non-additive Predictive Models." arXiv:1903.11420,

Model Inspection Problem to Visualization

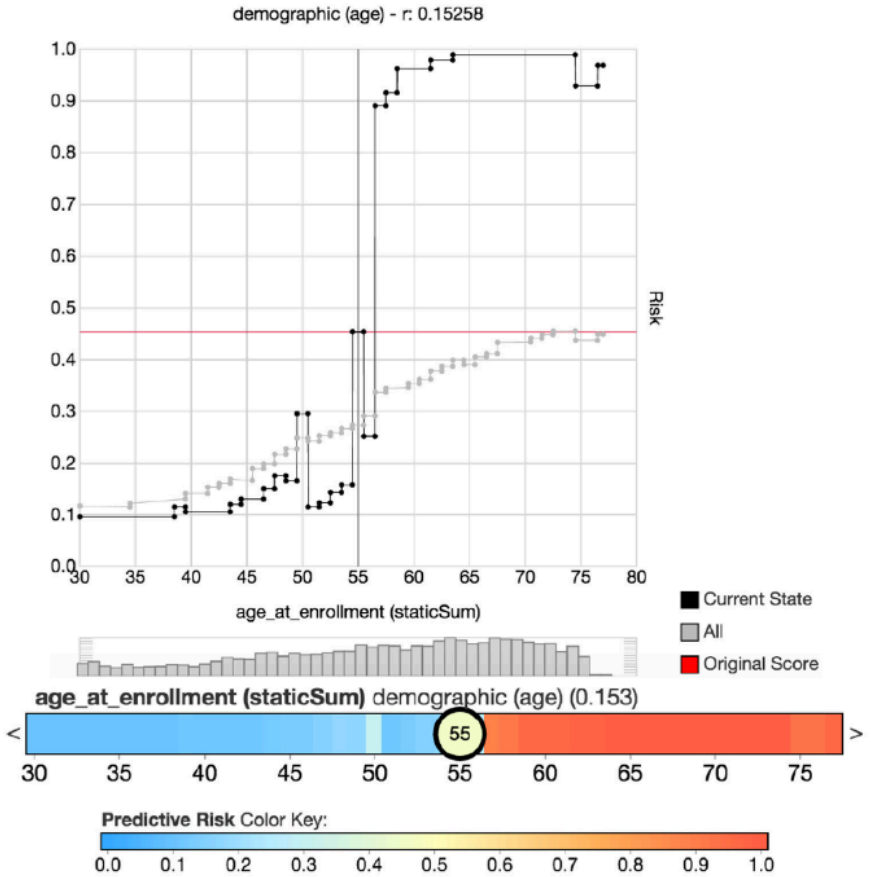
Provide a representation (visual or textual) for understanding either how the black box model works or why the black box returns certain predictions more likely than others.



Sensitivity Measures

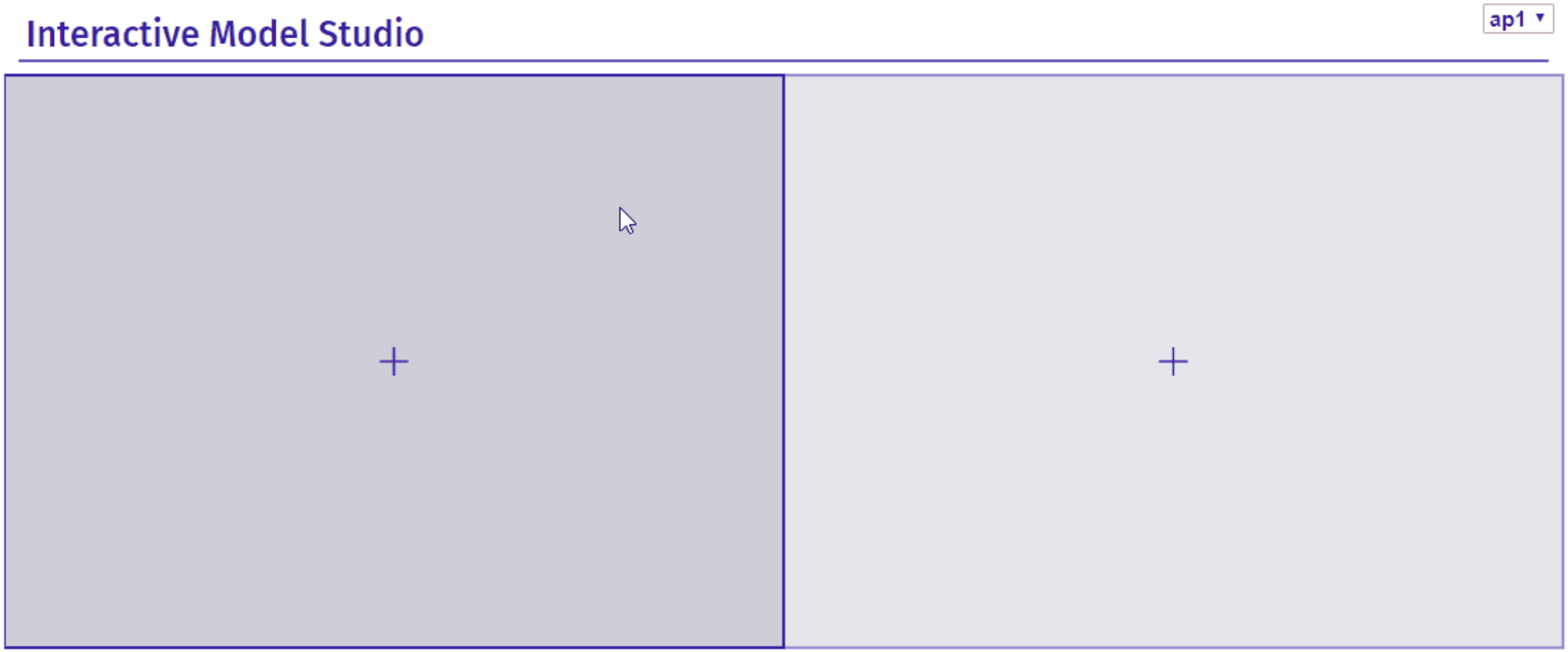


Paulo Cortez and Mark J. Embrechts. 2011. ***Opening black box data mining models using sensitivity analysis***. CIDM.



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

Linked display for property browsing



https://github.com/ModelOriented/modelStudio

Summary

- Efficient explanation through visual techniques
- Visualization strategies to reduce complexity of model visualization and multi-dimensionality
  - Visual clues
  - Interaction vs complexity
  - Linked displays for multi-dimensional exploration

Conclusions

Take Home Message

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, social sciences (multiple definitions)
- In Machine Learning:
  - Transparent design or post-hoc explanation?
  - Background knowledge matters!
  - We can scale-up symbolic reasoning by coupling it with representation learning on graphs.
- In AI (in general): many interesting / complementary approaches

Open The Black Box!

- ***To empower*** individual against undesired effects of automated decision making
- ***To reveal*** and protect new vulnerabilities
- ***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 align*** algorithms with human values
- ***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
- Is it possible to join ***local*** explanations to build a ***globally*** interpretable model?
- 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

XAI

European Research Council
Established by the European Commission

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

SoBigData

AI4EU

HUMANE AI

Track & Know

PRO-RES

Thank you!

Anna Monreale
University of Pisa

Salvo Rinzivillo
ISTI-CNR
Pisa

Riccardo Guidotti
ISTI-CNR
Pisa

Pasquale Minervini
University College
London