AAAI-19

On Explainable AI:

From Theory to Motivation, Applications and Limitations

Luca Costabello
Accenture Labs
@lukostaz

Fosca Giannotti
ISTI-CNR,
University of Pisa

Riccardo Guidotti
ISTI-CNR, University of Pisa
@rikdrive8s

Pascal Hitzler
Wright State University
@pascalhitzler

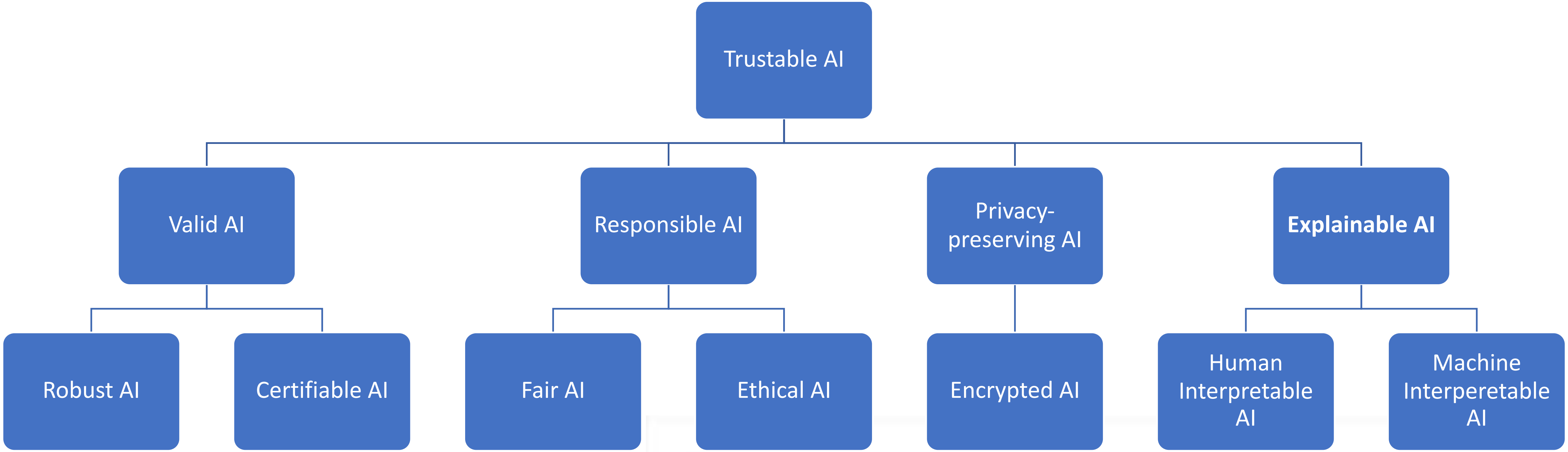
Freddy Lécué
Inria, France
CortAIx@Thales, Canada
@freddylecue

Pasquale Minervini
University College London
@PMinervini

Kamruzzaman Sarker
Wright State University
@smkpallob

xaitutorial2019.github.io

*AI Context for Industrial Adoption



- Trustable AI
  - Valid AI
    - Robust AI
    - Certifiable AI
  - Responsible AI
    - Fair AI
    - Ethical AI
  - Privacy-preserving AI
    - Encrypted AI
  - **Explainable AI**
    - Human Interpretable AI
    - Machine Interperetable AI

Disclaimer

- **As MANY interpretations as research areas** (check out work in Machine Learning vs Reasoning community)

- Not an exhaustive survey! Focus is on some promising approaches
- Massive body of literature (growing in time)
- Multi-disciplinary (AI – all areas, HCI, social sciences)
- Many domain-specific works hard to uncover
- Many papers do not include the keywords explainability/interpretability!

Explanation in AI

Explanation in AI aims to create a suite of techniques that produce more explainable models, while maintaining a high level of searching, learning, planning, reasoning performance: optimization, accuracy, precision; and enable human users to understand, appropriately trust, and effectively manage the emerging generation of AI systems .

Motivation (1)

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

Opinion

The New York Times

OP-ED CONTRIBUTOR

When a Computer Program Keeps You in Jail

By Rebecca Wexler

June 13, 2017

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

How We Analyzed the COMPAS Recidivism Algorithm

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

May 23, 2016

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

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

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

[Rudin 2018]

Motivation (2)

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

The Big Read Artificial intelligence + Add to myFT

Insurance: Robots learn the business of covering risk

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

Oliver Ralph MAY 16, 2017

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

Explainable Machine Learning Challenge

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

Motivation (3)

- Healthcare
  - Applying ML methods in medical care is problematic.
  - AI as 3rd-party actor in physician-patient relationship
  - Responsibility, confidentiality?
  - Learning must be done with available data.

    Cannot randomize cares given to patients!
  - Must validate models before use.

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

Researchers say use of artificial intelligence in medicine raises ethical questions

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

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

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

Rich Caruana
Microsoft Research
rcaruana@microsoft.com

Yin Lou
LinkedIn Corporation
ylou@linkedin.com

Johannes Gehrke
Microsoft
johannes@microsoft.com

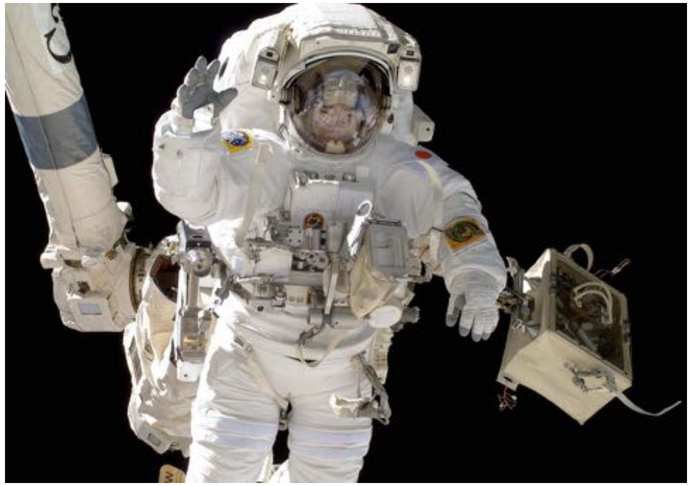
Paul Koch
Microsoft Research
paulkoch@microsoft.com

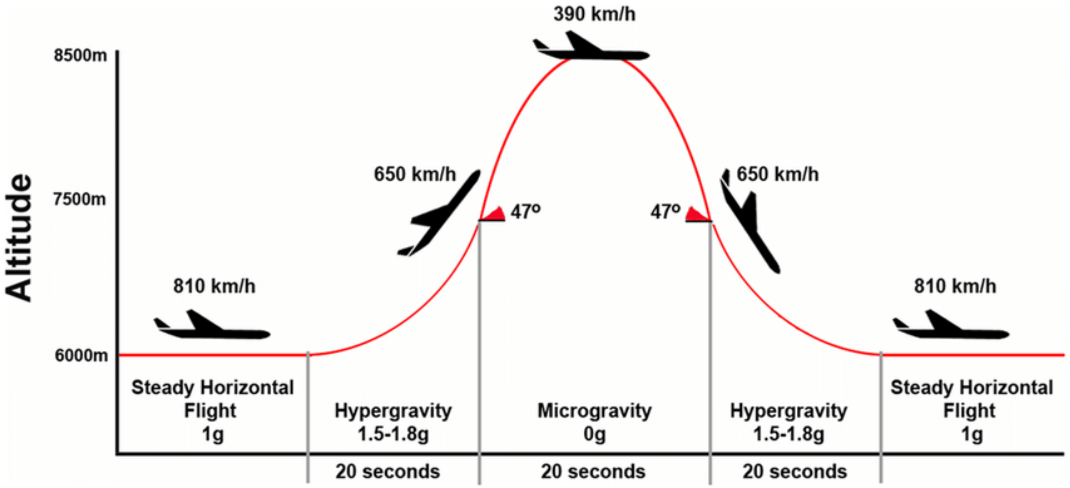
Marc Sturm
NewYork-Presbyterian Hospital
mas9161@nyp.org

Noémie Elhadad
Columbia University
noemie.elhadad@columbia.edu

Motivation (4)

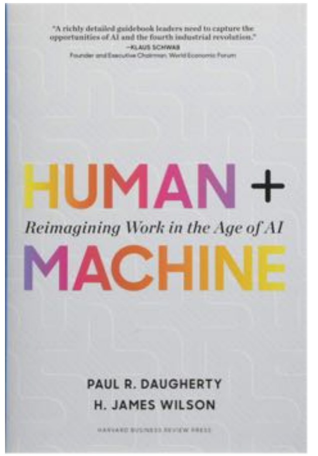
- Critical Systems





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

The Need for Explanation



HUMAN +
Reimagining Work in the Age of AI
MACHINE

PAUL R. DAUGHERTY
H. JAMES WILSON

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

[Lepri et al. 2018]

Tutorial Outline (1)

- **Explanation in AI** 8:40 – 9:30
  - Definitions & Properties
  - Explanations in different AI fields
  - The Role of Humans
  - Evaluation Protocols & Metrics
- **Explainable Machine Learning** 9:30 – 10:30
  - What is a Black Box?
  - Interpretable, Explainable, and Comprehensible Models
  - Open the Black Box Problems
- **Break** 10:30 – 11:00

Tutorial Outline (2)

- **Explainable AI with Background Knowledge** **11:00 – 11:15**
  - Explainability in terms of Domain Knowledge
  - State of the art to use domain knowledge
- **Machine Learning on Knowledge Graphs** **11:15 – 12:00**
  - Knowledge Graphs
  - Relational Learning
  - Neuro-Symbolic Reasoning and Neural Theorem Provers
- **Applications** **12:00 – 12:30**

References

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

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

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

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

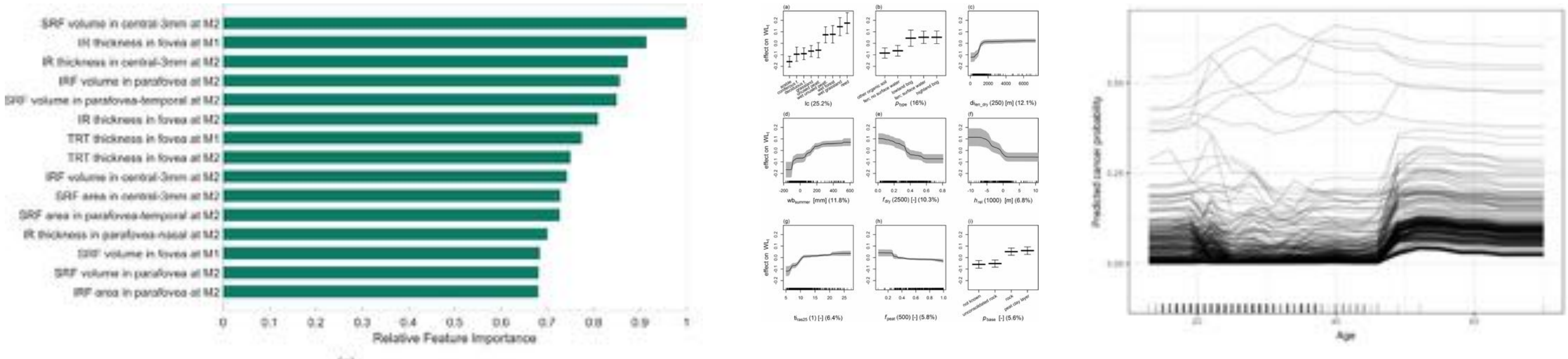
Explanation in AI

Luca Costabello
Accenture Labs, Ireland
@lukostaz

Freddy Lécué
Inria, France
CortAIx@Thales, Canada
@freddylecue

Overview of explanation in different AI fields (1)

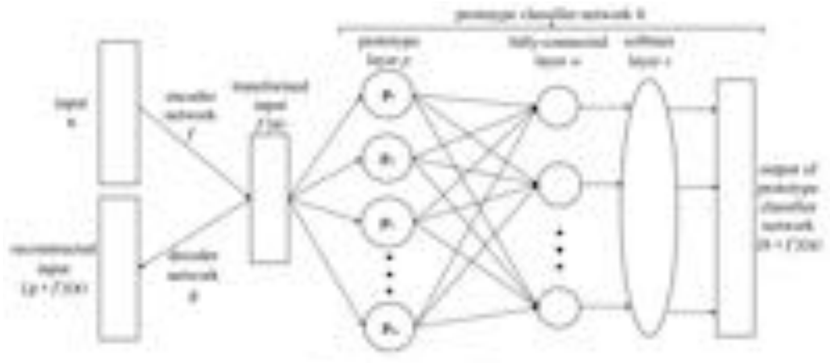
- Machine Learning



Feature Importance, Partial Dependence Plot, Individual Conditional Expectation

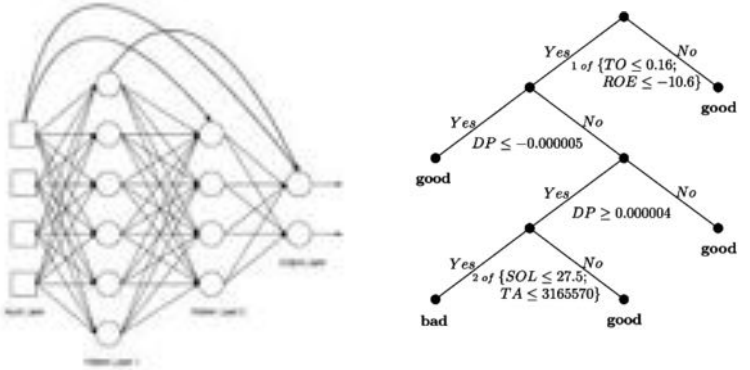
Interpretable Models:

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



Auto-encoder

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

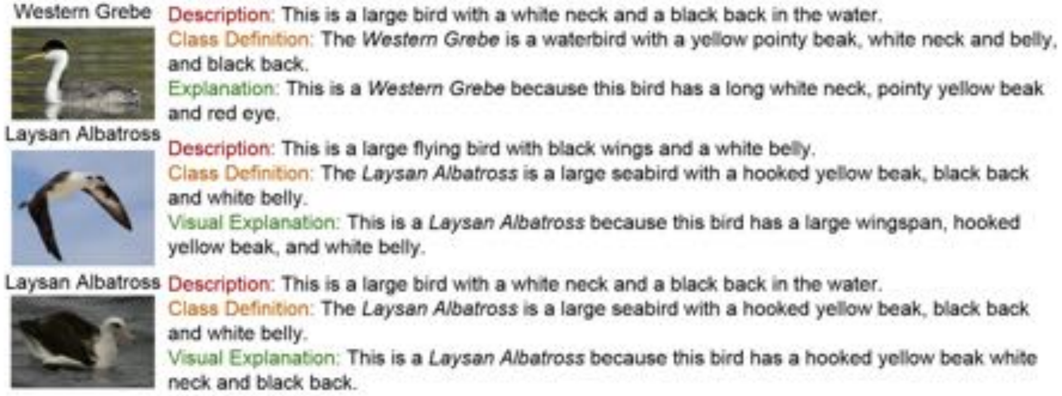


Surogate Model

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

Overview of explanation in different AI fields (2)

- Computer Vision



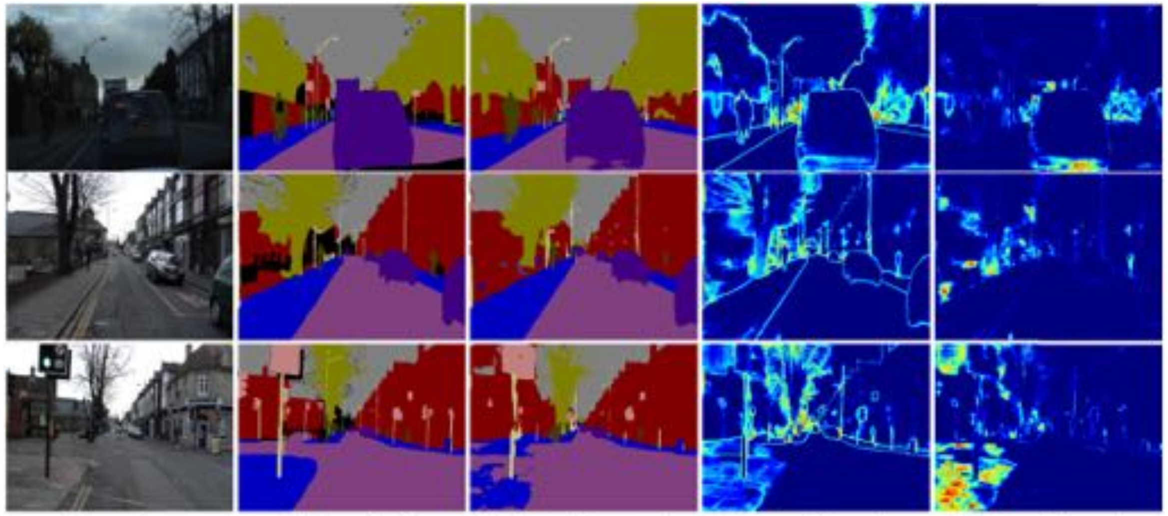
Description: This is a large bird with a white neck and a black back in the water.
Class Definition: The *Western Grebe* is a waterbird with a yellow pointy beak, white neck and belly, and black back.
Explanation: This is a *Western Grebe* because this bird has a long white neck, pointy yellow beak and red eye.

Description: This is a large flying bird with black wings and a white belly.
Class Definition: The *Laysan Albatross* is a large seabird with a hooked yellow beak, black back and white belly.
Visual Explanation: This is a *Laysan Albatross* because this bird has a large wingspan, hooked yellow beak, and white belly.

Description: This is a large bird with a white neck and a black back in the water.
Class Definition: The *Laysan Albatross* is a large seabird with a hooked yellow beak, black back and white belly.
Visual Explanation: This is a *Laysan Albatross* because this bird has a hooked yellow beak white neck and black back.

Visual Explanation

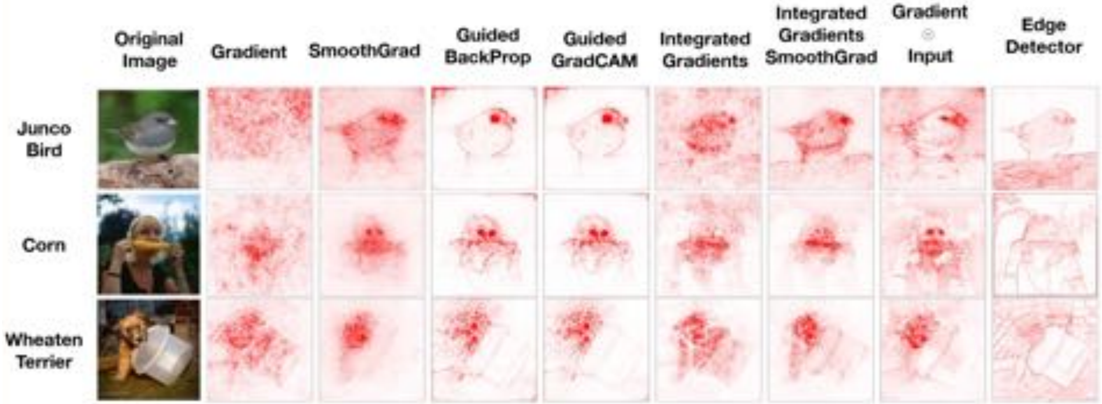
Lisa Anne Hendricks, Zeynep Akata, Marcus Rohrbach, Jeff Donahue, Bernt Schiele, Trevor Darrell: Generating Visual Explanations. ECCV (4) 2016: 3-19



(a) Input Image (b) Ground Truth (c) Semantic Segmentation (d) Aleatoric Uncertainty (e) Epistemic Uncertainty

Uncertainty Map

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

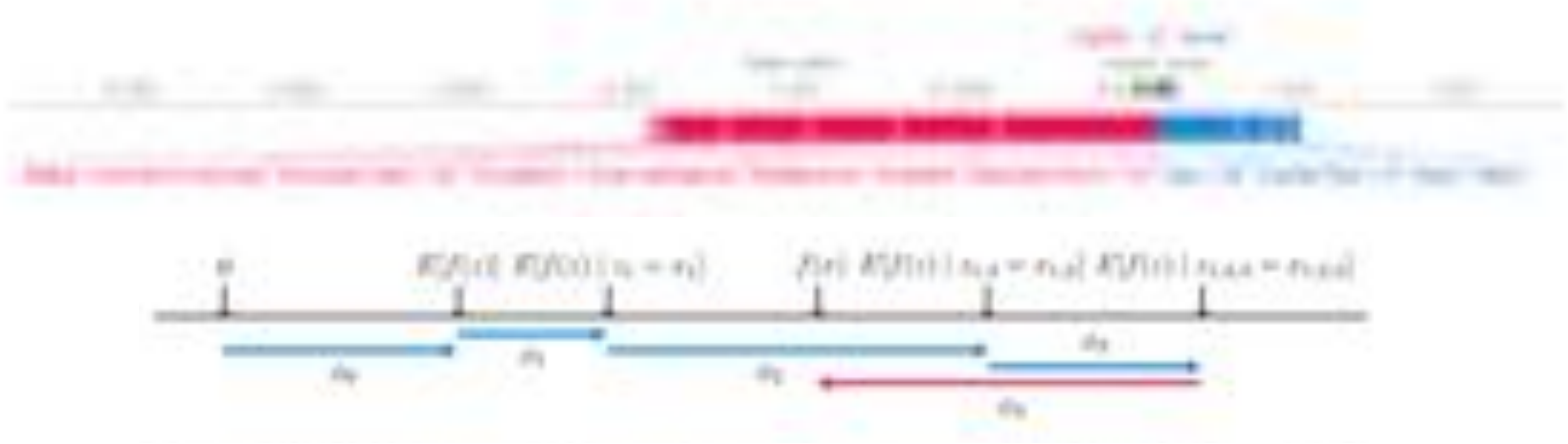


Saliency Map

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

Overview of explanation in different AI fields (3)

- Game Theory

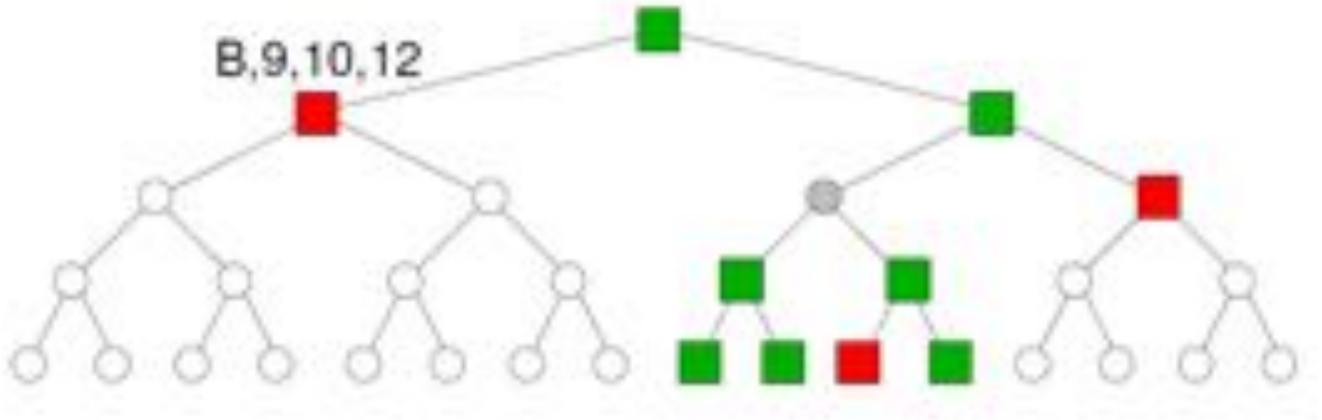


Shapley Additive Explanation

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

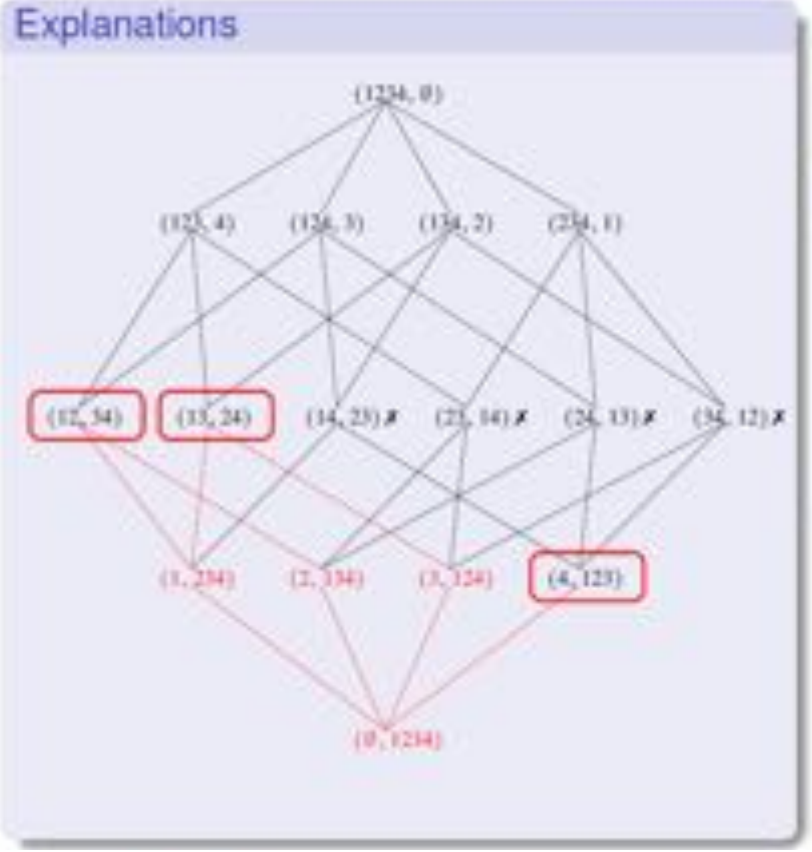
Overview of explanation in different AI fields (4)

- Search and Constraint Satisfaction



Conflicts resolution

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



Constraints relaxation

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

Overview of explanation in different AI fields (5)

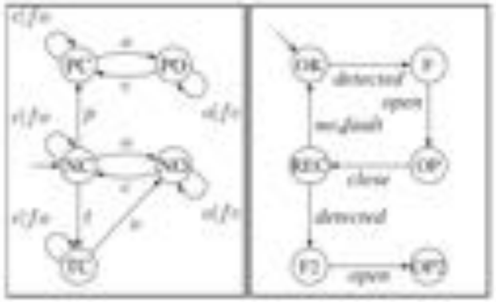
- Knowledge Representation and Reasoning

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

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

Abduction Reasoning (in Bayesian Network)

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



Diagnosis Inference

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

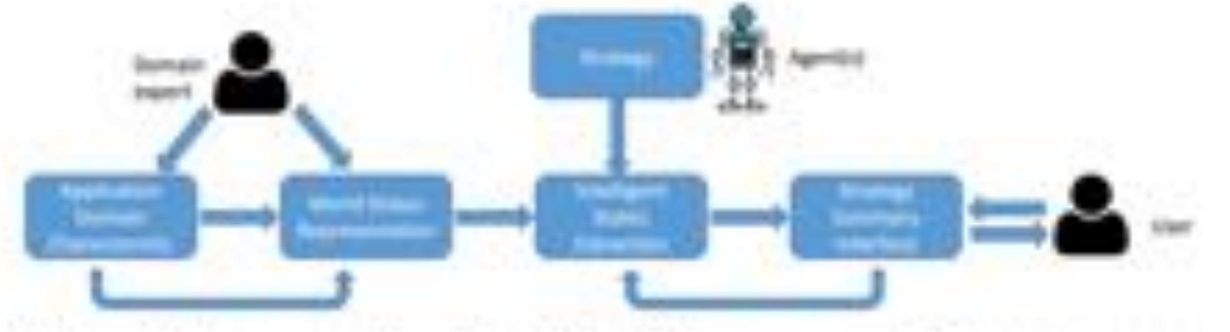
Overview of explanation in different AI fields (6)

- Multi-agent Systems

MAS INFRASTRUCTURE	INDIVIDUAL AGENT INFRASTRUCTURE
MAS INTEROPERATION Translation Services Interoperation Services	**INTEROPERATION** Interoperation Modules
CAPABILITY TO AGENT MAPPING Middle Agents	**CAPABILITY TO AGENT MAPPING** Middle Agents Components
NAME TO LOCATION MAPPING ANS	**NAME TO LOCATION MAPPING** ANS Component
SECURITY Certificate Authority Cryptographic Services	**SECURITY** Security Module private/public Keys
PERFORMANCE SERVICES MAS Monitoring Reputation Services	**PERFORMANCE SERVICES** Performance Services Modules
MULTIAGENT MANAGEMENT SERVICES Logging, Activity Visualization, Launching	**MANAGEMENT SERVICES** Logging and Visualization Components
ACL INFRASTRUCTURE Public Ontology Protocols Servers	**ACL INFRASTRUCTURE** ACL Parser Private Ontology Protocol Engine
COMMUNICATION INFRASTRUCTURE Discovery Message Transfer	**COMMUNICATION MODULES** Discovery Component Message Transfer Module
OPERATING ENVIRONMENT Machines, OS, Network Multicast Transport Layer: TCP/IP, Wireless, Infrared, SSL	

Explanation of Agent Conflicts and Harmful Interactions

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



Agent Strategy Summarization

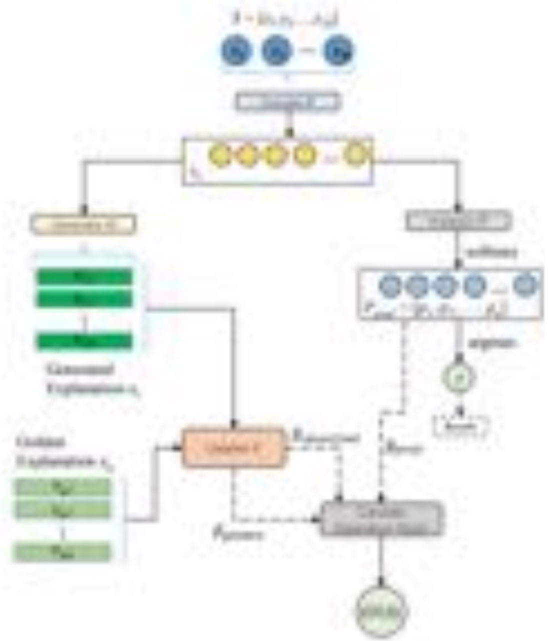
Ofra Amir, Finale Doshi-Velez, David Sarne: Agent Strategy Summarization. AAMAS 2018: 1203-1207

Explainable Agents

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

Overview of explanation in different AI fields (7)

- NLP



Fine-grained explanations are in the form of:

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

Explainable NLP

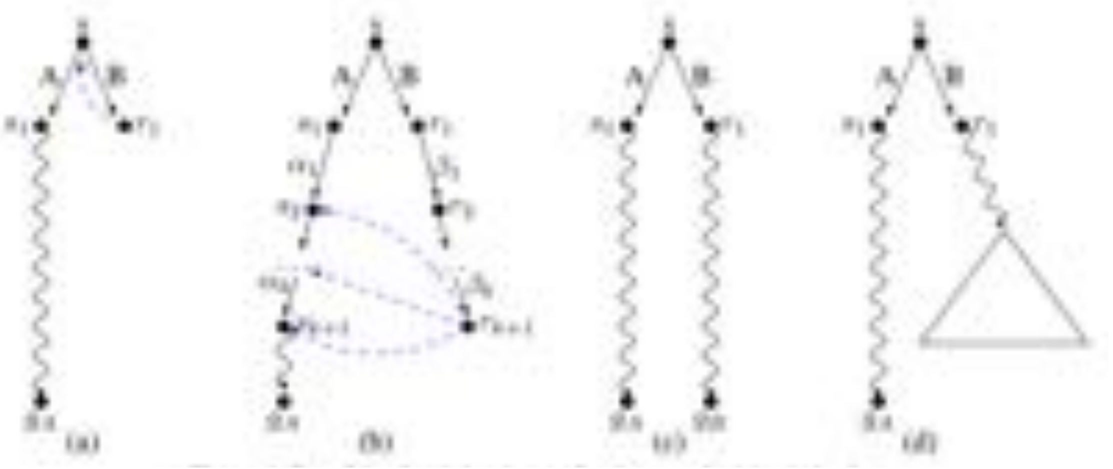
Hui Liu, Qingyu Yin, William Yang Wang: Towards Explainable NLP: A Generative Explanation Framework for Text Classification. CoRR abs/1811.00196 (2018)

LIME for NLP

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

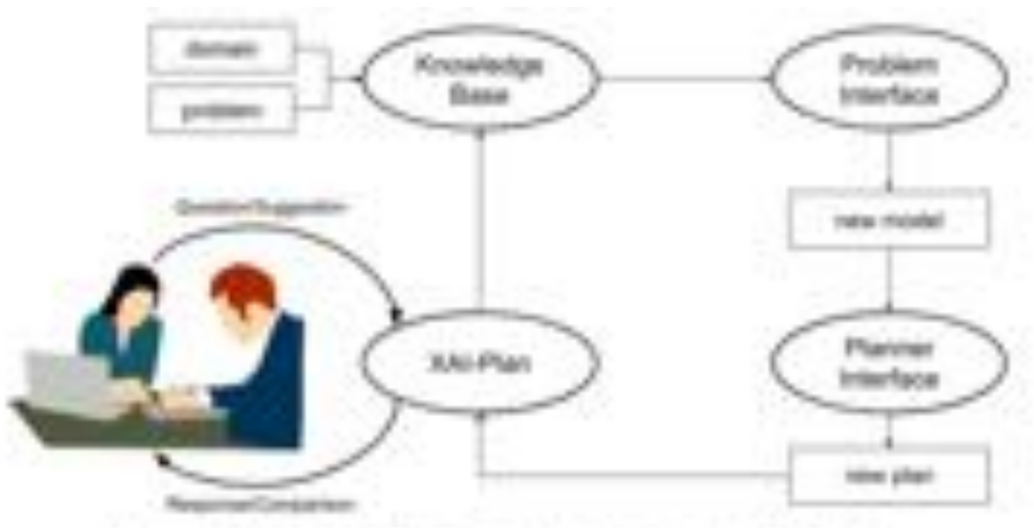
Overview of explanation in different AI fields (8)

- Planning and Scheduling



Human-in-the-loop Planning

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

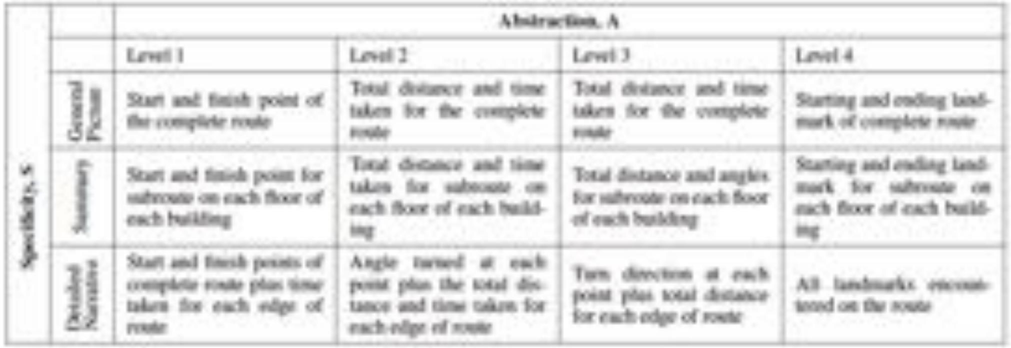


XAI Plan

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

Overview of explanation in different AI fields (9)

- Robotics



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

Narration of Autonomous Robot Experience

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

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

Robot: I have decided to turn left.

Human: Why did you do that?

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

Human: How confident are you in this decision?

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

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

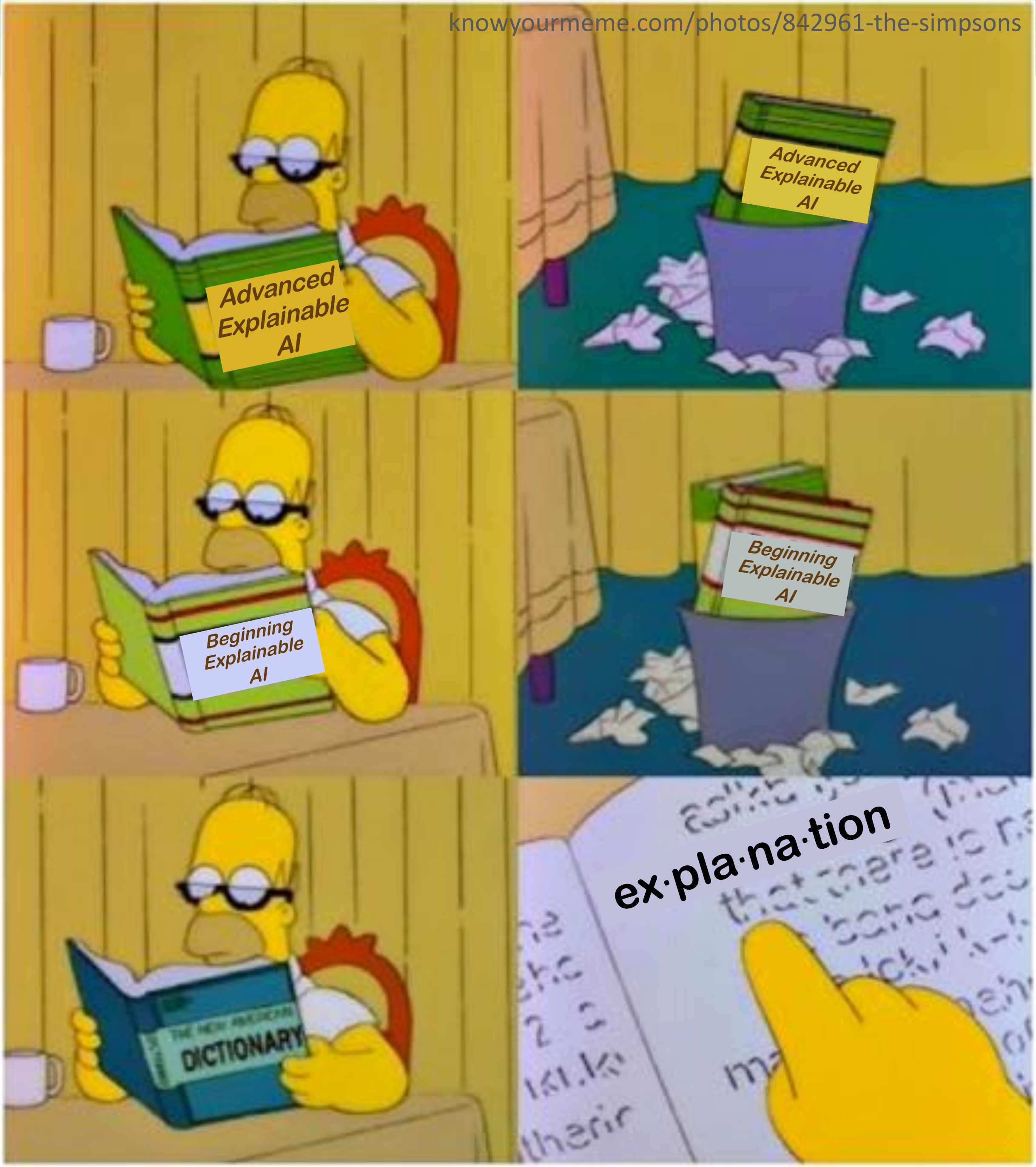
Robot: Here is the histogram of all training examples that reached this leaf. 80% of examples where this area was above 20 cm predicted the appropriate action to be "drive forward".

From Decision Tree to human-friendly information

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

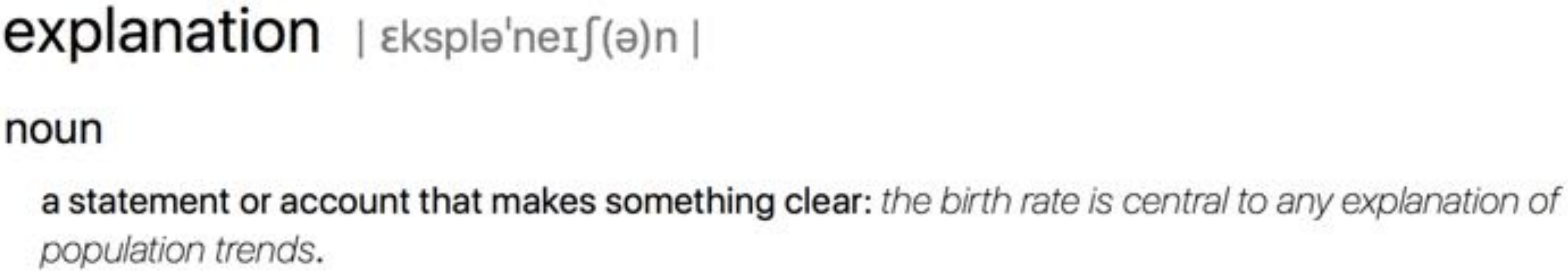
The Need to Explain

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



knowyourmeme.com/photos/842961-the-simpsons

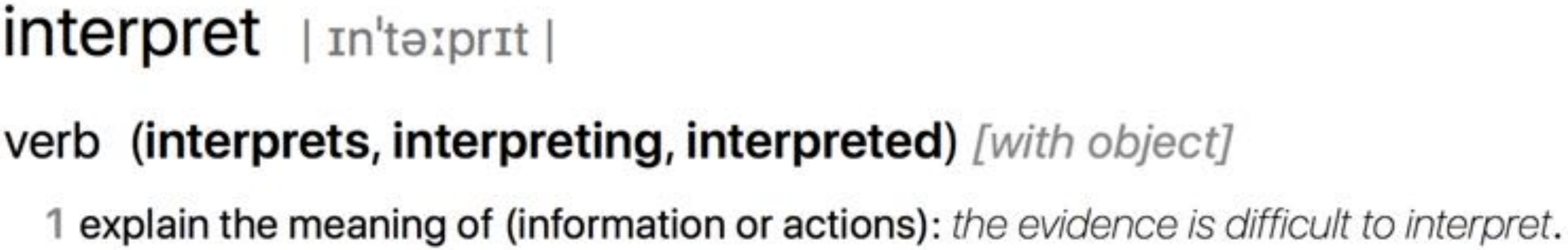
Oxford Dictionary of English



explanation | ɛkspləˈneɪʃ(ə)n |

noun

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

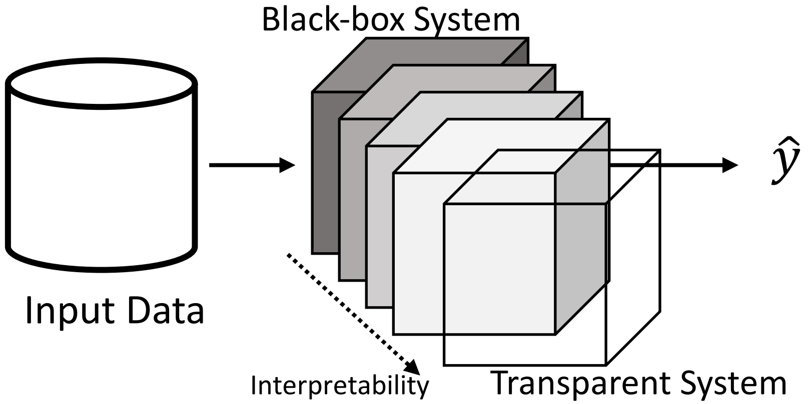


interpret | ɪnˈtəːprɪt |

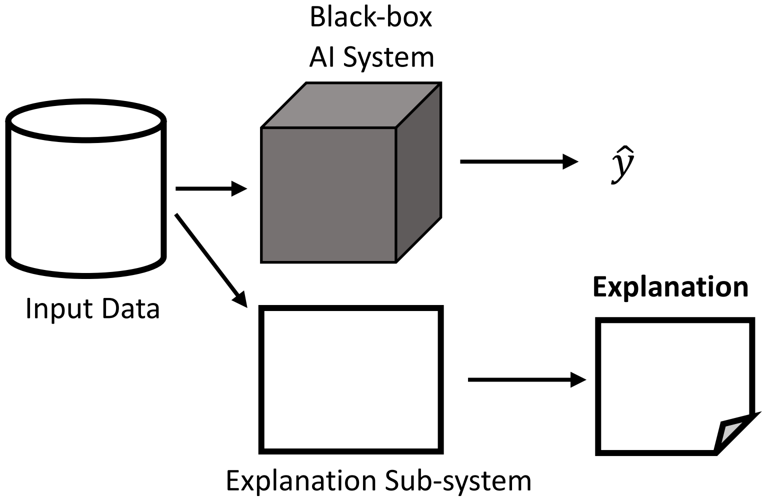
verb **(interprets, interpreting, interpreted)** *[with object]*

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

Transparent Design vs Post-hoc Explanation



Transparent design reveals *how* a model functions.



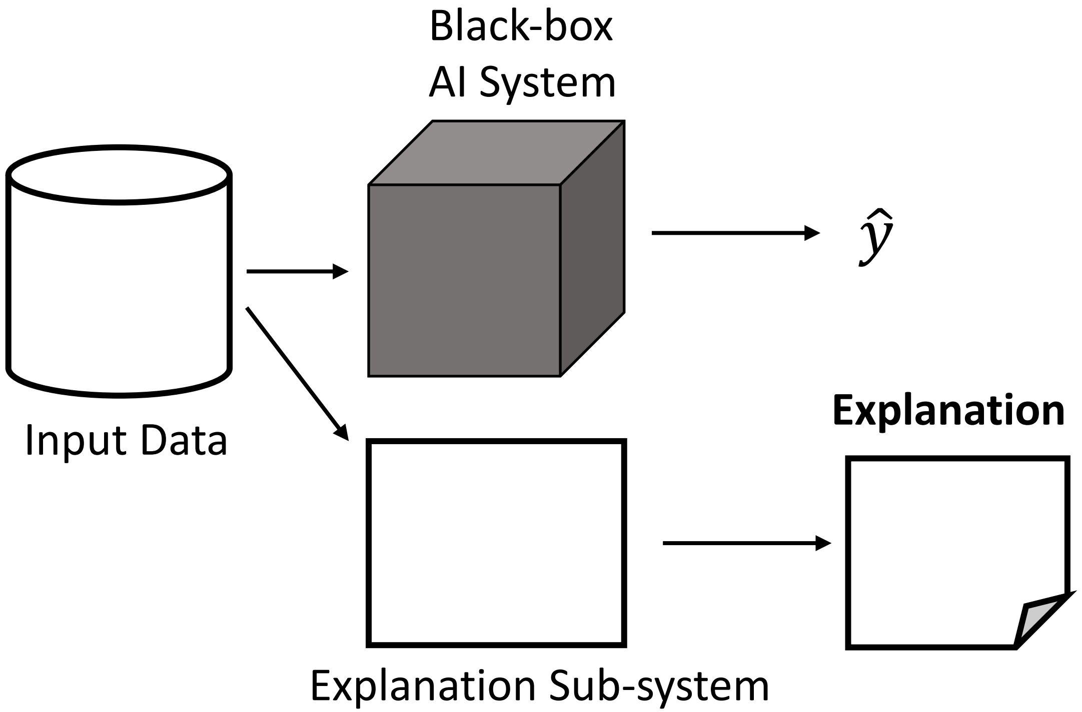
Post-hoc Explanation explains *why* a black-box model behaved that way.

[Mittelstadt et al. 2018]

So, What is an Explanation?

- **No formal, technical, agreed upon definition!**
- Comprehensive philosophical overview out of scope of the tutorial [Miller 2017]
- Not limited to machine learning!

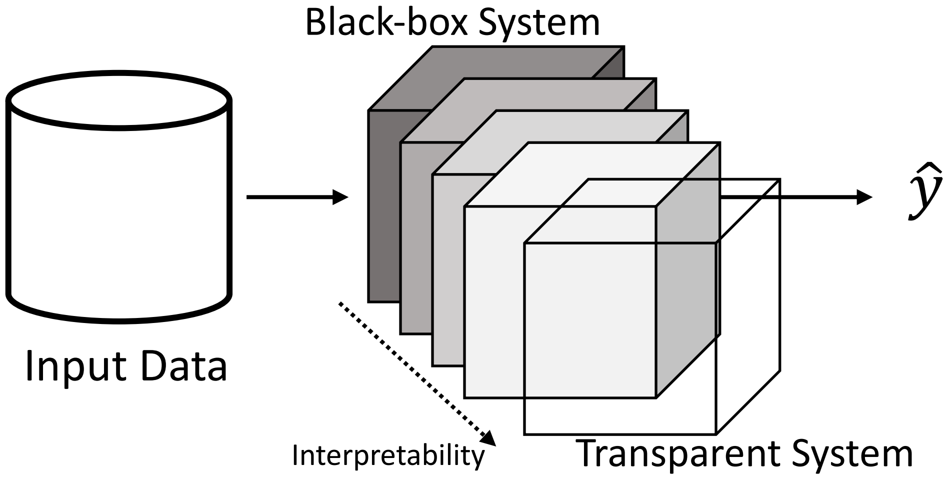
[Lipton 2016, Tomsett et al. 2018, Rudin 2018]



What About Interpretability?

- Interpretability as Multi-Faceted Concept
  - Interpretability is an ill-defined term!
  - **Not** a monolithic concept

[Lipton 2016]



Levels of Model Transparency

Simulatability

Understanding of the functioning of the **model**

- Can a human *easily* predict outputs?
- Can a human examine the model all at once?

Transparent model

Decomposability

Understanding at the level of **single components** (e.g. parameters)

Transparent Model Components

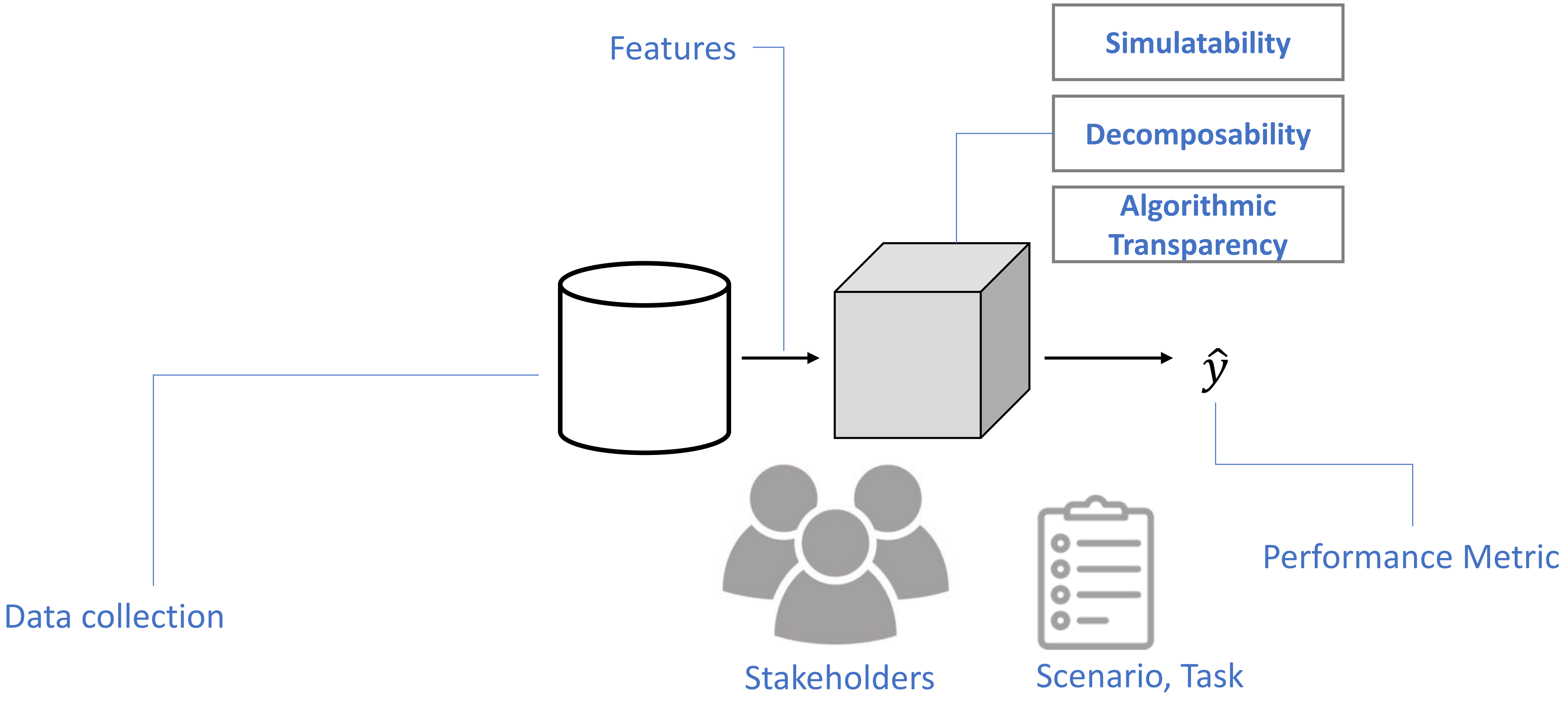
Algorithmic Transparency

Understanding at the level of **training algorithm**

Transparent Training Algorithm

[Lipton 2016, Lepri et al. 2017, Mittelstadt et al. 2018, Weld and Bansal 2018]

Interpretability Goes Beyond the Model



Features

Simulatability

Decomposability

Algorithmic Transparency

$\hat{y}$

Data collection

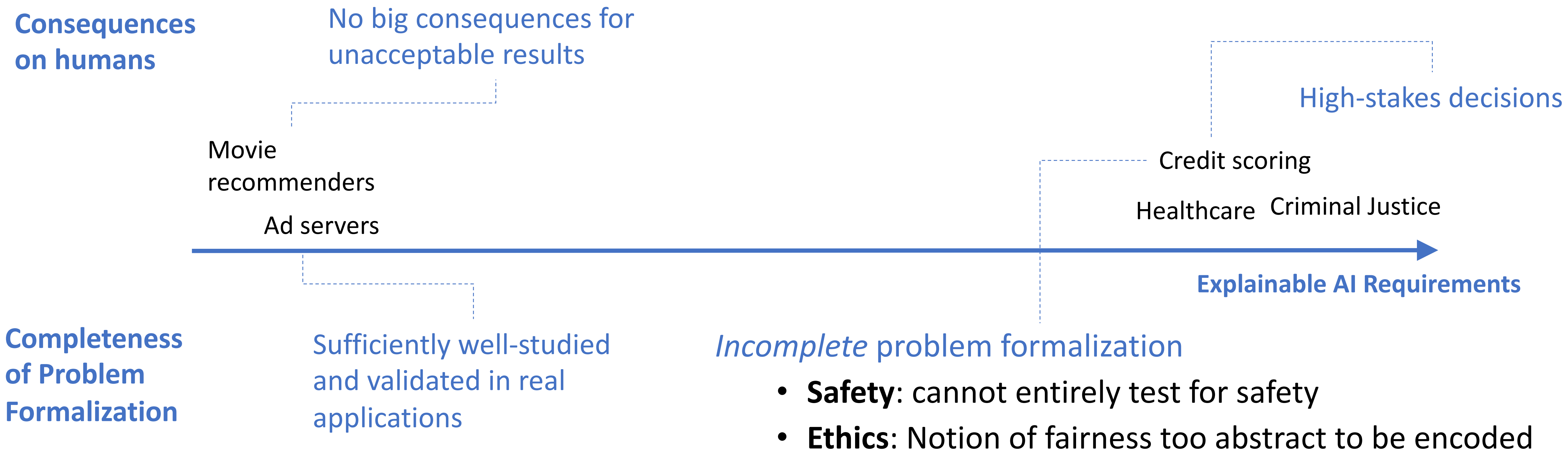
Stakeholders

Scenario, Task

Performance Metric

Desire for Explainable AI **Must be Justified**

Interpretability comes at cost: Trade-off interpretability/predictive power



Consequences on humans

No big consequences for unacceptable results

Movie recommenders

Ad servers

High-stakes decisions

Credit scoring

Healthcare

Criminal Justice

Explainable AI Requirements

Completeness of Problem Formalization

Sufficiently well-studied and validated in real applications

Incomplete problem formalization

- **Safety**: cannot entirely test for safety
- **Ethics**: Notion of fairness too abstract to be encoded

[Freitas 2014 , Lipton 2016, Doshi-velez and Kim 2017, Wend and Bansal 2017 , Rudin 2018]

High-Stakes Scenarios Deserve Transparent Models

- Post-hoc explanations can be unreliable
- Design white-box, interpretable models straight away!
- (Or retro-fit approximate but interpretable models over complex ones)
- Problem: with thousands+ features DNNs perform better: post-hoc explanation the only way (?)

[Rudin 2018, Mittelstadt et al. 2018]

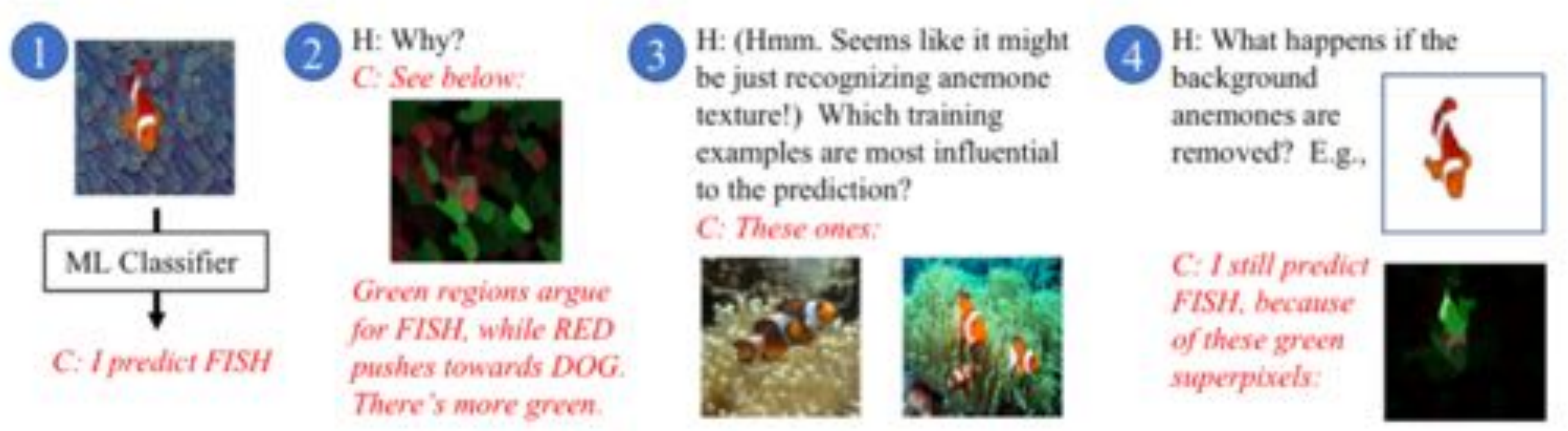
(Some) Desired Properties of Explainable AI Systems

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

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

Explanation as *System-Human Conversation*

[Weld and Bansal 2018]



① ML Classifier

C: I predict FISH

② H: Why?

C: See below:

Green regions argue for FISH, while RED pushes towards DOG. There's more green.

③ H: (Hmm. Seems like it might be just recognizing anemone texture!) Which training examples are most influential to the prediction?

C: These ones:

④ H: What happens if the background anemones are removed? E.g.,

C: I still predict FISH, because of these green superpixels:

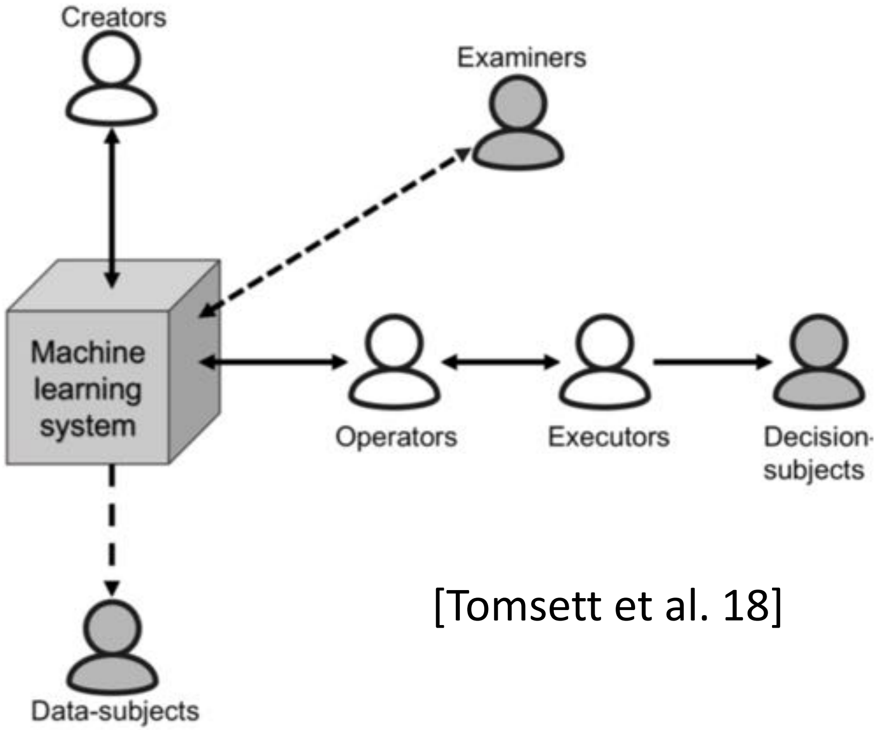
- Humans may have follow-up questions
- Explanations cannot answer all users' concerns

Role-based Interpretability

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

No Universally Interpretable Model!

- **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?"
- **C-suite**



[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]

Designing Explanations is Task-Related

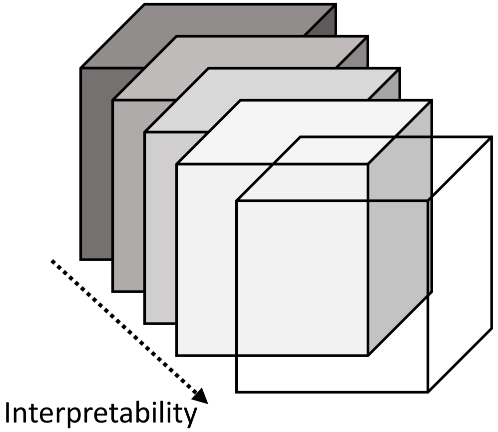
- Interpretability is always scenario-dependent!

  W*hat does interpretability mean* in a specific context? Ask the experts!
- What is the ultimate goal of the explanation in that specific **context,** for that specific **task**?
- How incomplete is the problem formulation?
- Time constraints
- Which user expertise?

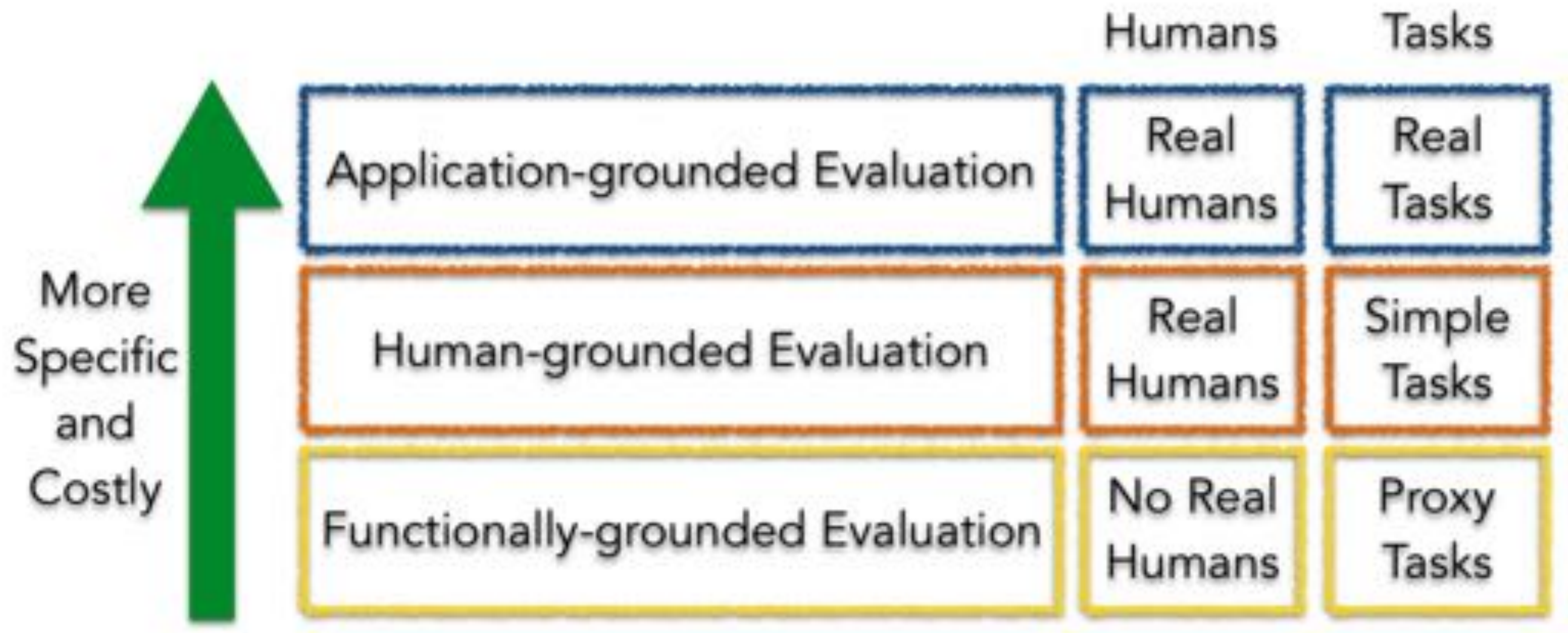
[Lipton 2016, Rudin 2018, Doshi-Velez and Kim 2017]

Evaluation: Interpretability as Latent Property

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



Evaluation Approaches



More Specific and Costly

	Humans	Tasks
Application-grounded Evaluation	Real Humans	Real Tasks
Human-grounded Evaluation	Real Humans	Simple Tasks
Functionally-grounded Evaluation	No Real Humans	Proxy Tasks

[Doshi-Velez and Kim 2017]

Human-Independent Metrics: Size

- Size is over-simplistic [Freitas 14]
  - E.g.: # nodes in a decision tree, size of a local explanation
  - Humans can handle at most 7±2 symbols [Miller1956, Rudin2018]
  - Size does not capture *semantics* of the model
  - Extreme simplicity insufficient! e.g. medical experts and larger models, [Freitas 2014]
  - What does *too large* mean?

[Doshi-Velez and Kim 2017, Poursabzi-Sangdeh 18]

Human-based Evaluation is Essential

Evaluation criteria for Explanations [Miller, 2017]

- Truth & probability
- Usefulness, relevance
- Coherence with prior belief
- Generalization

Cognitive chunks = basic explanation units (for different explanation needs)

- Which basic units for explanations?
- How many?
- How to compose them?
- Uncertainty & end users?

[Doshi-Velez and Kim 2017, Poursabzi-Sangdeh 18]

Open Challenges

- More formal studies on interpretability
- *Rigorous, agreed upon* evaluation protocols
- More work on transparent design
- Human involvement (e.g. better interactive, “social” explanations) [Miller 2017]
- Define industry standards (e.g. AI Service Factsheet [Hind et al. 2018)]
- Improve existing legislation
  - “Right to explanation” vs “right to be informed” [Wachter et al. 2017]
  - Legislation & Explanations: How accurate ? How complete? How faithful to the model? [Rudin 2018]

tl;dr

- Explanations and interpretability are required for better human trust, system debug, and legal compliance.
- No monolithic, agreed upon definition of Explainable AI
- Adoption spans multiple AI fields
- Explainability, interpretability come at a cost
- Design with humans and task in mind
- Human-based evaluation is essential

References

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Explainable Machine Learning

Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri,

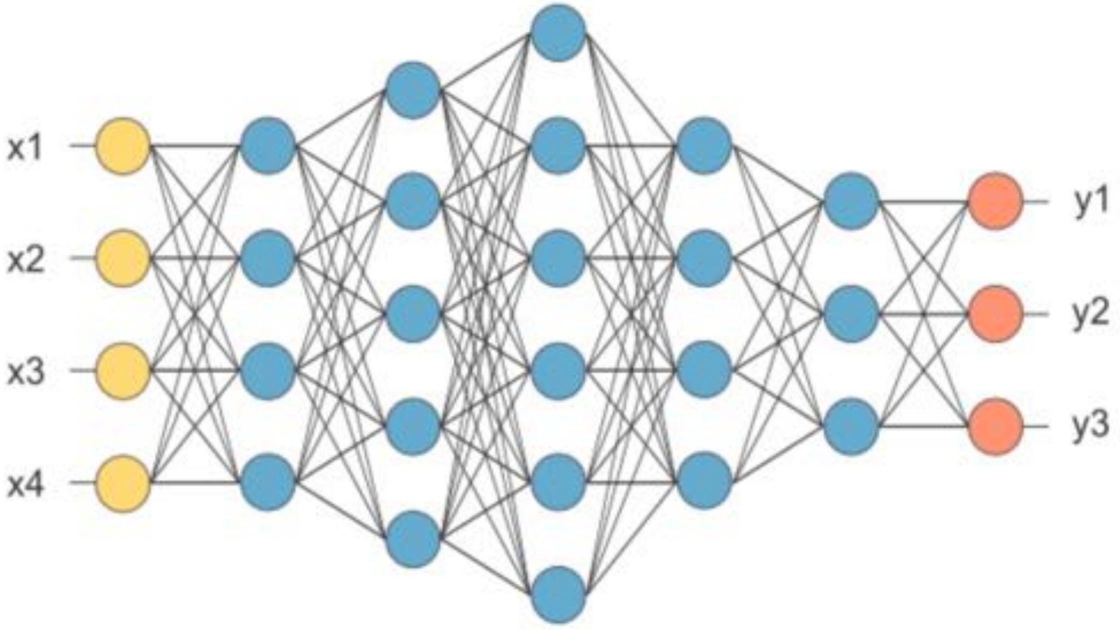
Franco Turini, Fosca Giannotti, ***Dino Pedreschi***

KDD Lab

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

UNIVERSITÀ DI PISA

Black Box Model



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

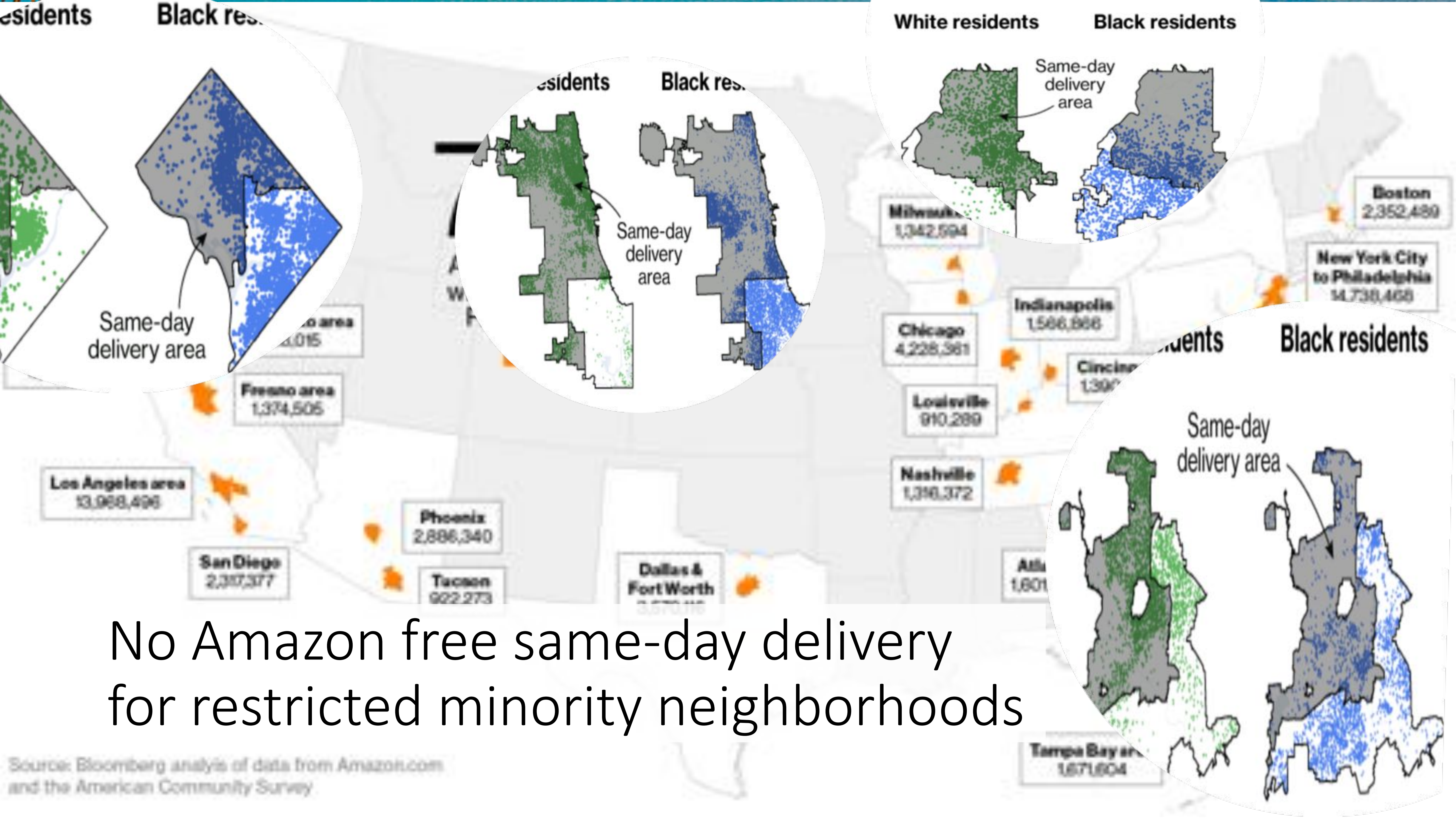
- Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). ***A survey of methods for explaining black box models***. *ACM Computing Surveys (CSUR)*, *51*(5), 93.

Needs For Interpretable Models

COMPAS recidivism black bias

DYLAN FUGETT	**BERNARD PARKER**
Prior Offense 1 attempted burglary	**Prior Offense** 1 resisting arrest without violence
Subsequent Offenses 3 drug possessions	**Subsequent Offenses** None
LOW RISK 3	HIGH RISK 10

Fugett was rated low risk after being arrested with cocaine and marijuana. He was arrested three times on drug charges after that.



No Amazon free same-day delivery for restricted minority neighborhoods

Source: Bloomberg analysis of data from Amazon.com and the American Community Survey

The background bias

(a) Husky classified as wolf

(b) Explanation

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.

Interpretable, Explainable and Comprehensible Models

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

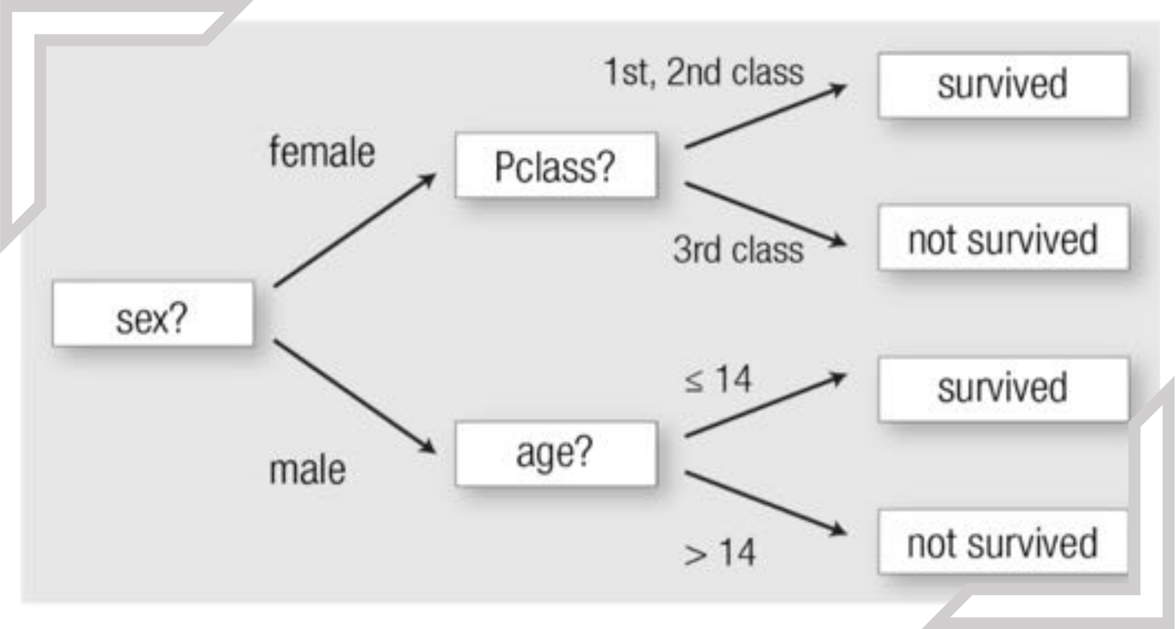
- ***Fairness***: the model guarantees the protection of groups against discrimination.
- ***Privacy***: the model does not reveal sensitive information about people.
- ***Respect Monotonicity***: the increase of the values of an attribute either increase or decrease in a monotonic way the probability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

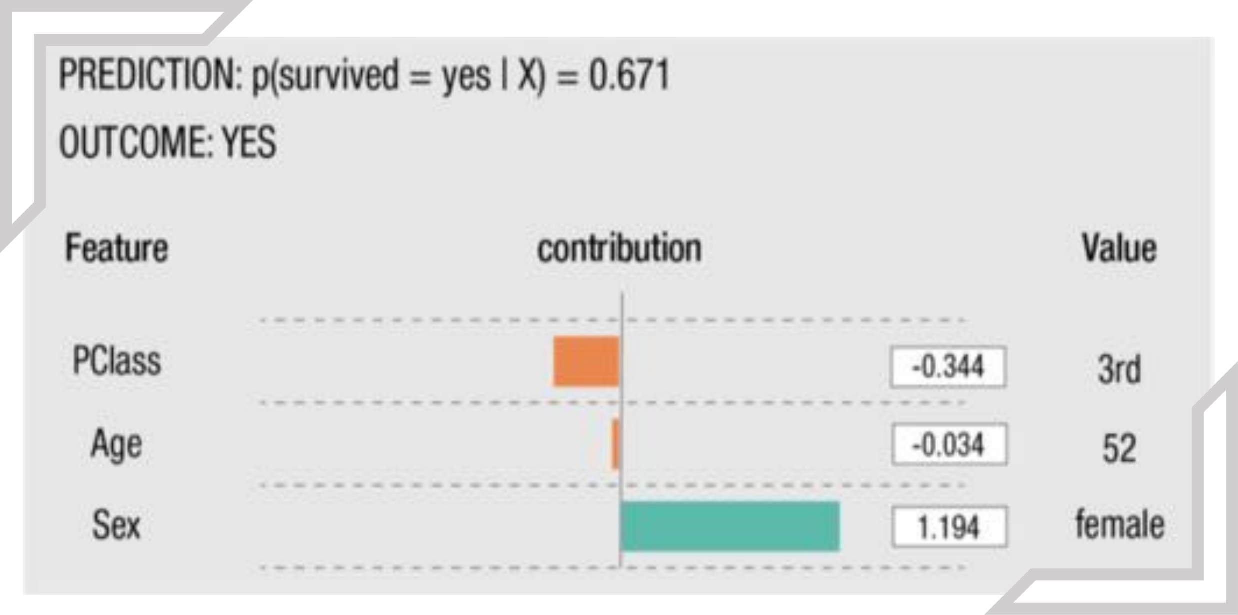
Desiderata of an Interpretable Model

- ***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

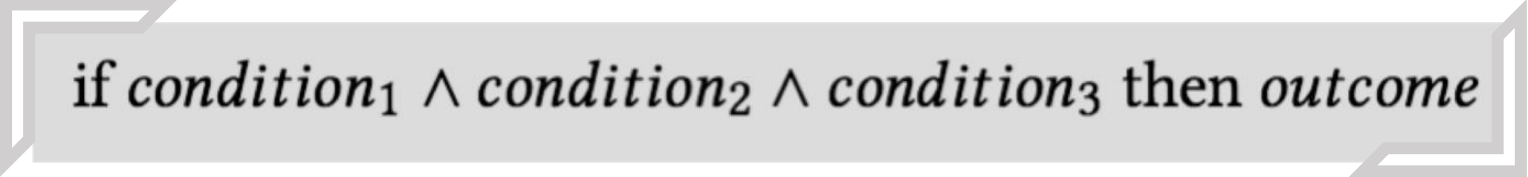


Decision Tree


PREDICTION: p(survived = yes | X) = 0.671

OUTCOME: YES

Linear Model



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

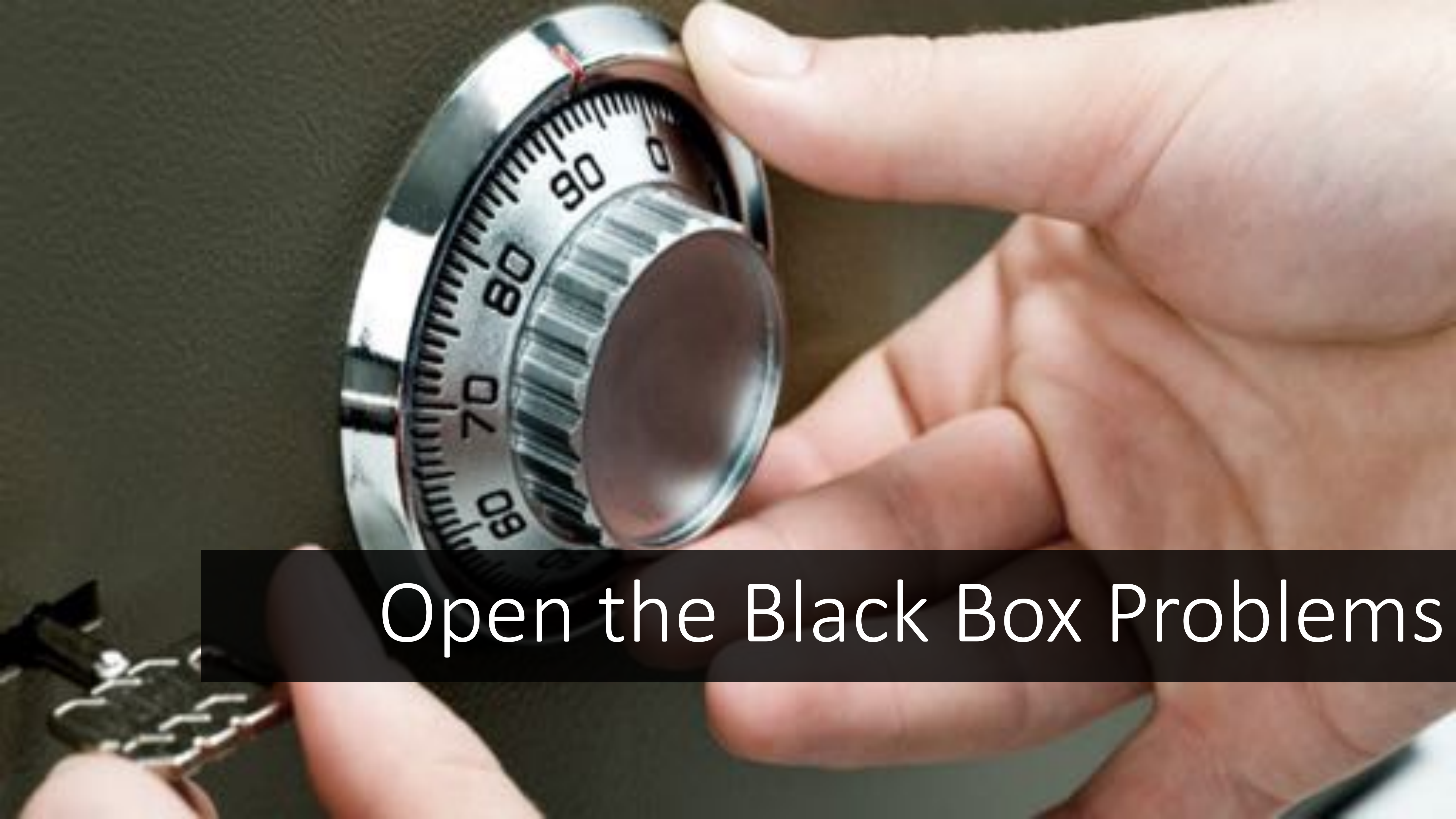
Rules

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.

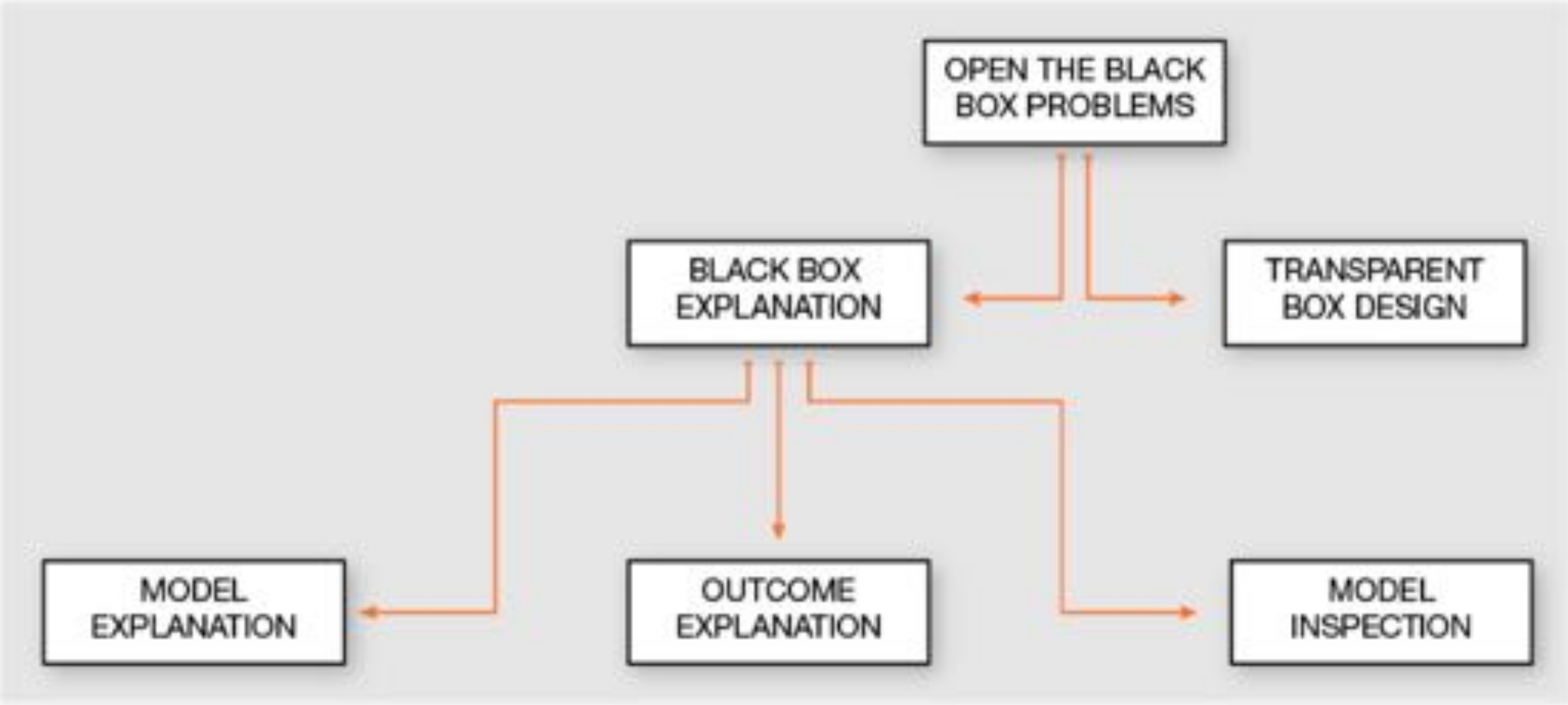
- Linear Model: number of non zero weights in the model.
- Rule: number of attribute-value pairs in condition.
- Decision Tree: estimating the complexity of a tree can be hard.

- 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



OPEN THE BLACK BOX PROBLEMS

BLACK BOX EXPLANATION

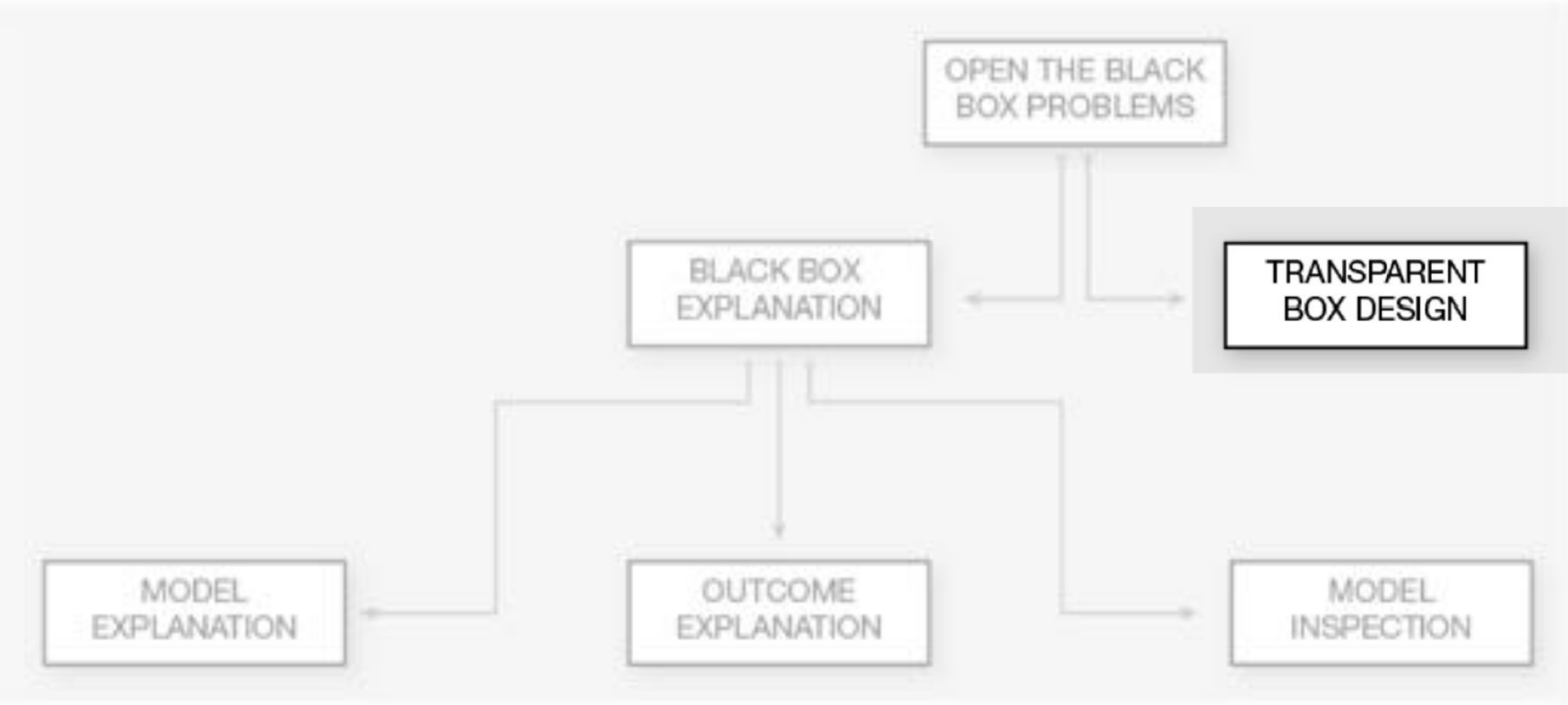
TRANSPARENT BOX DESIGN

MODEL EXPLANATION

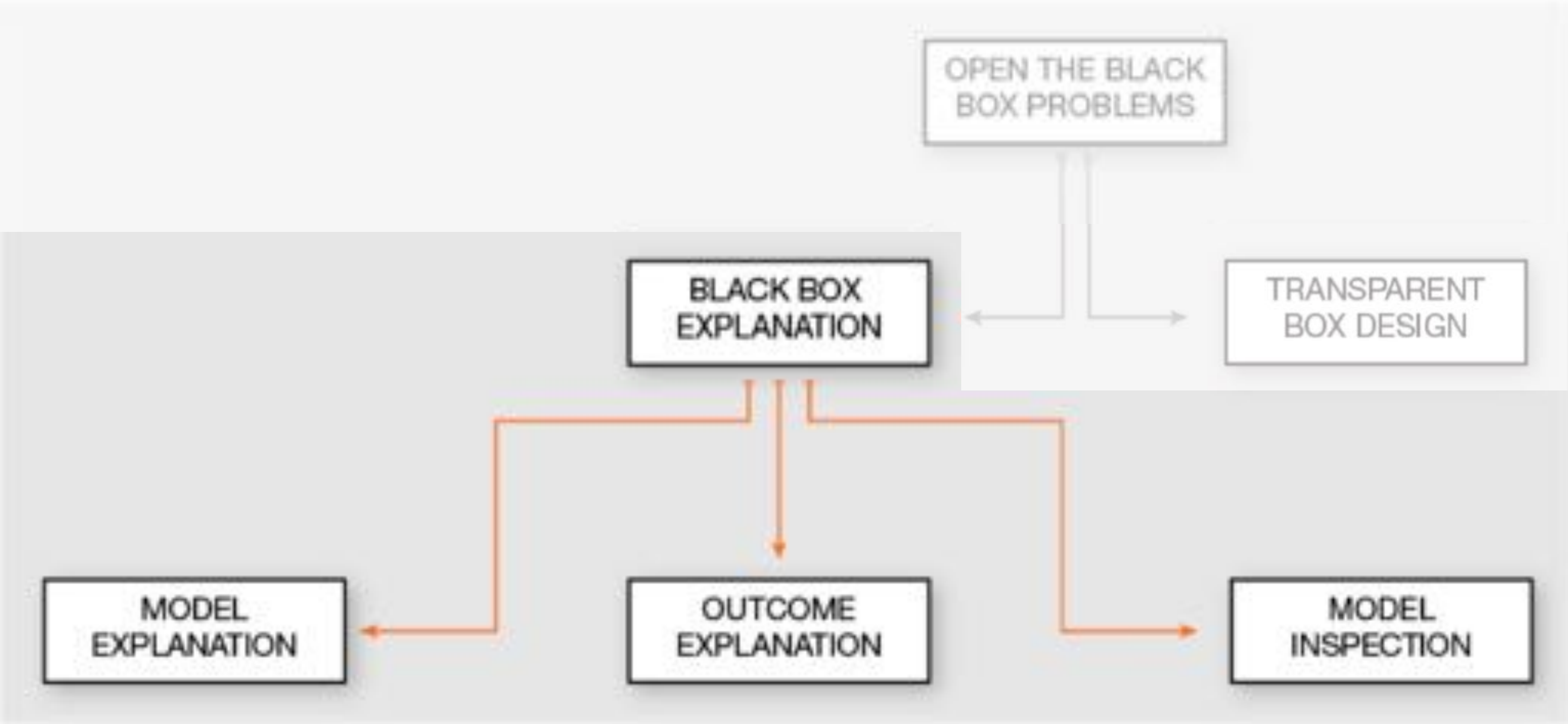
OUTCOME EXPLANATION

MODEL INSPECTION

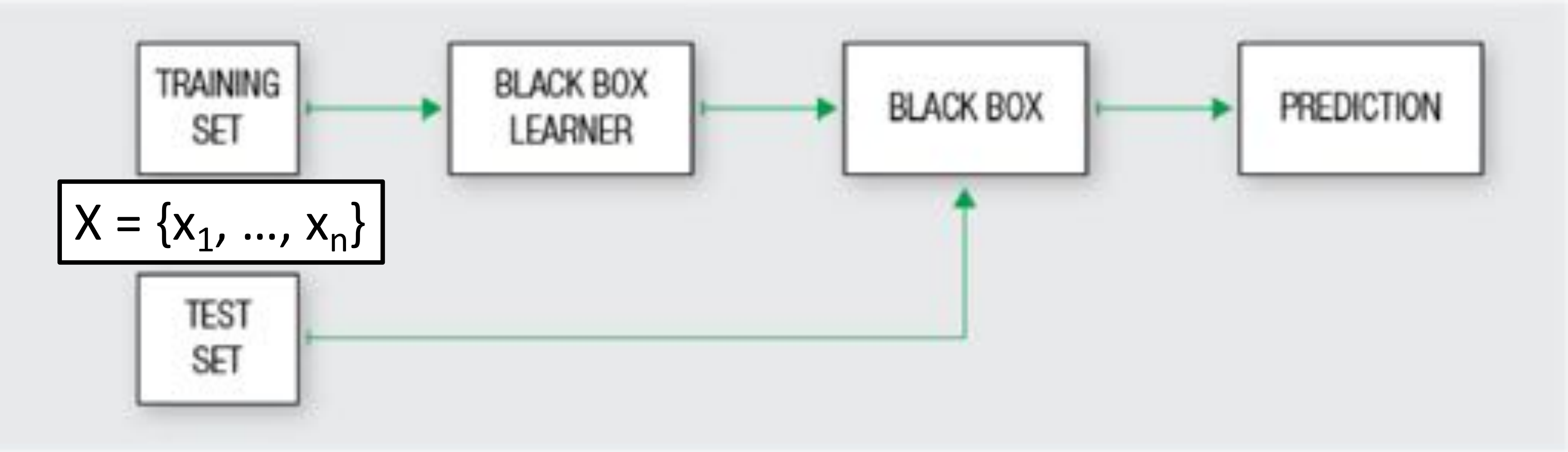
XbD – eXplanation by Design



BBX - Black Box eXplanation



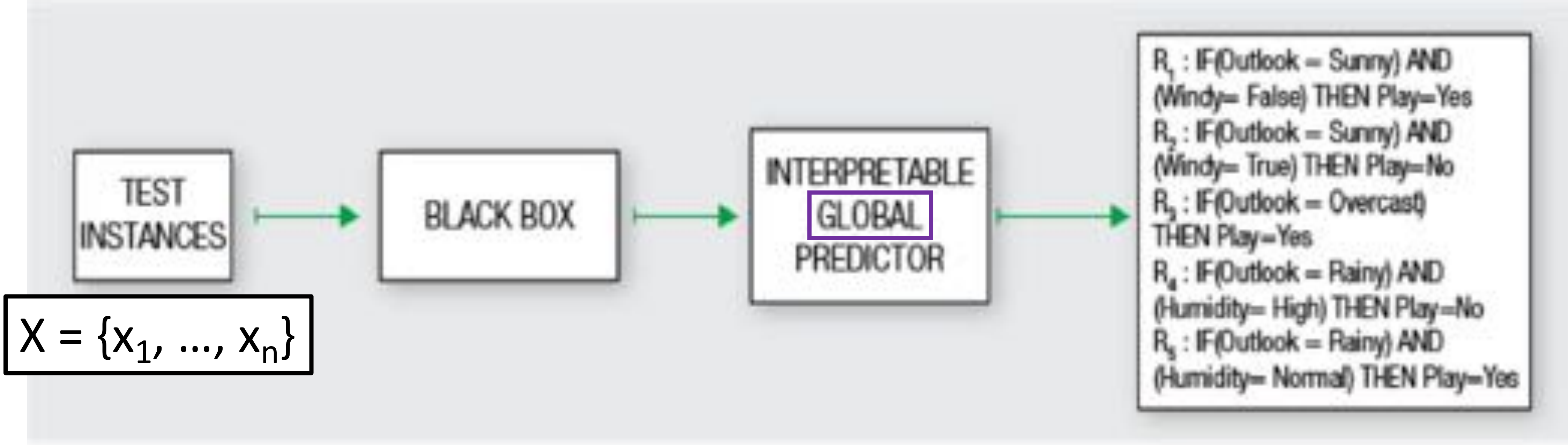
Classification Problem



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

Model Explanation Problem

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

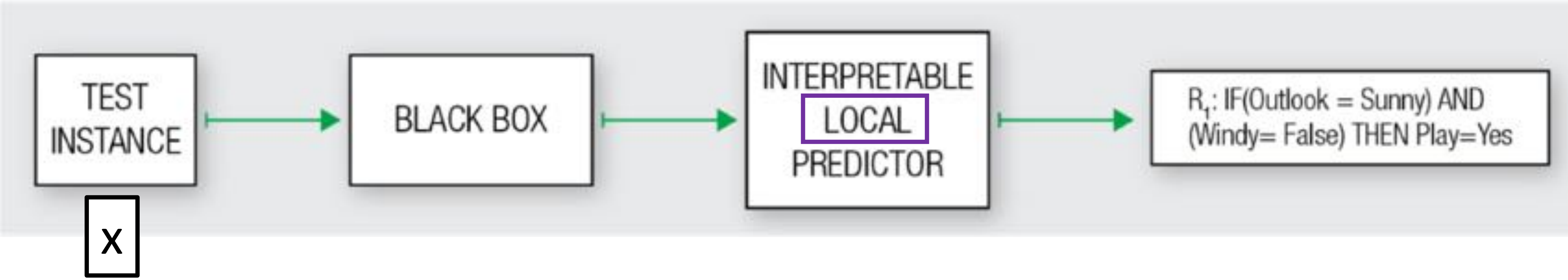


$X = \{x_1, \ldots, x_n\}$

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

Outcome Explanation Problem

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



TEST INSTANCE

x

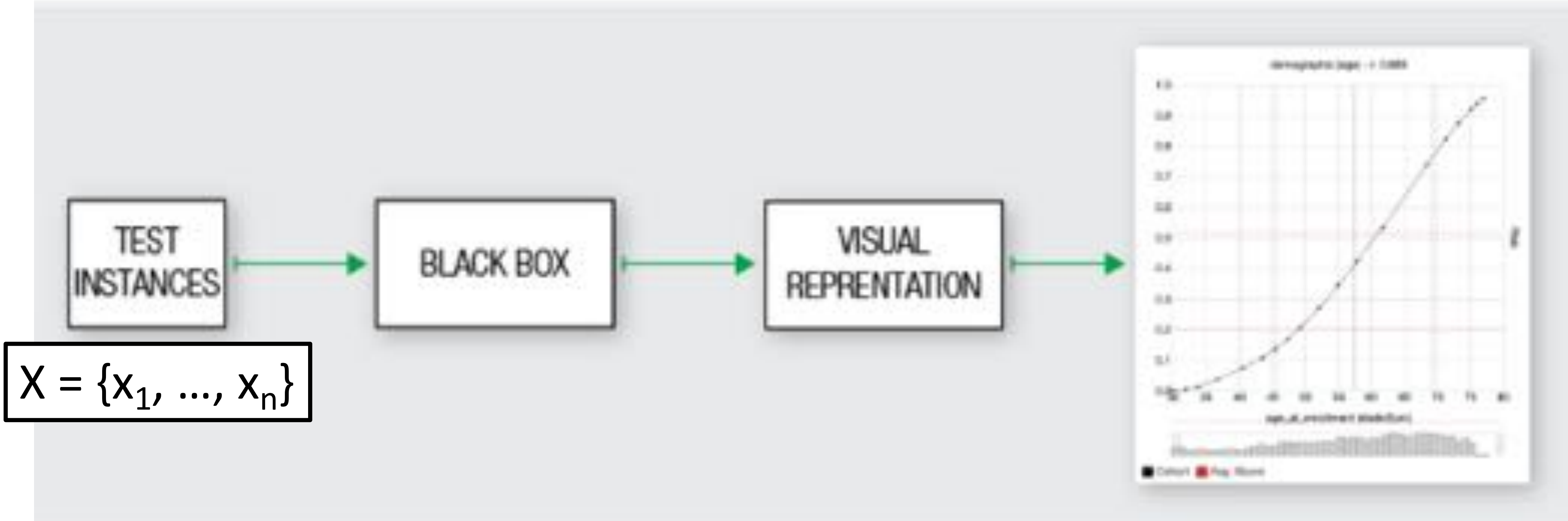
BLACK BOX

INTERPRETABLE LOCAL PREDICTOR

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

Model Inspection Problem

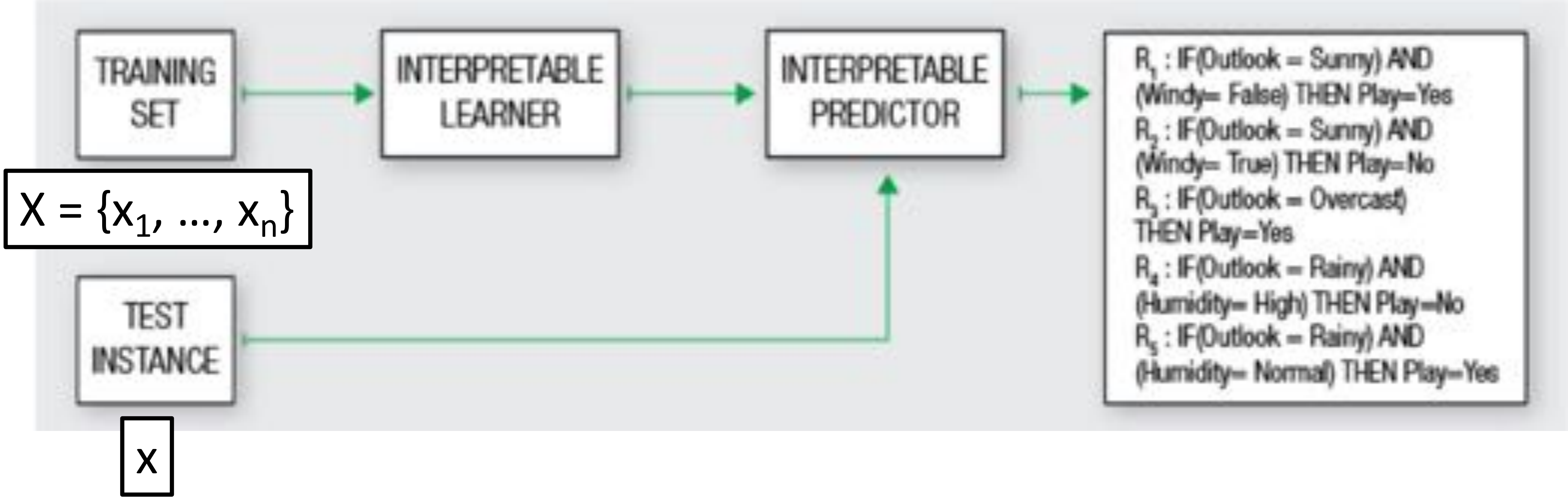
Provide a representation (visual or textual) for understanding either how the black box model works or why the black box returns certain predictions more likely than others.



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

Transparent Box Design Problem

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



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

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

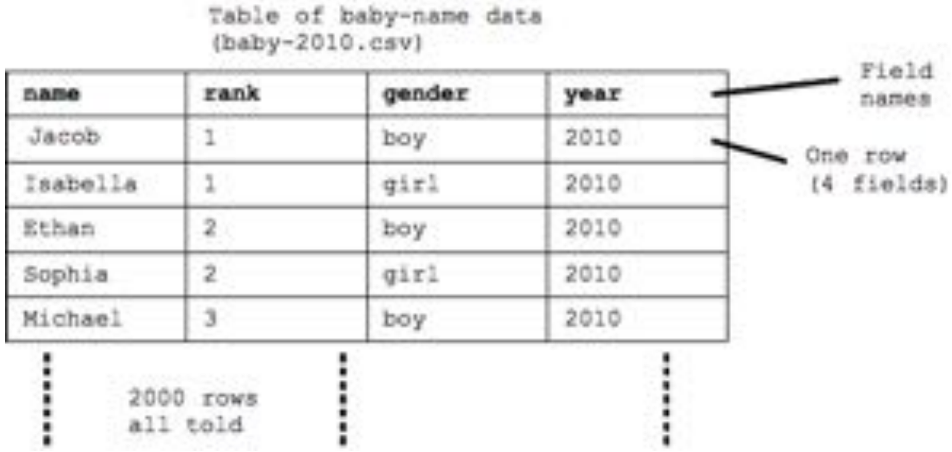


Table of baby-name data
(baby-2010.csv)

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**)

© Bernard Castelein / naturepl.com

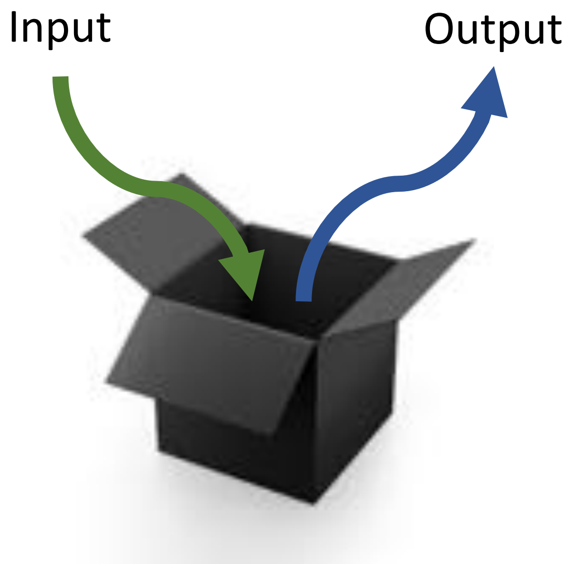
Text
(**TXT**)

Explanators

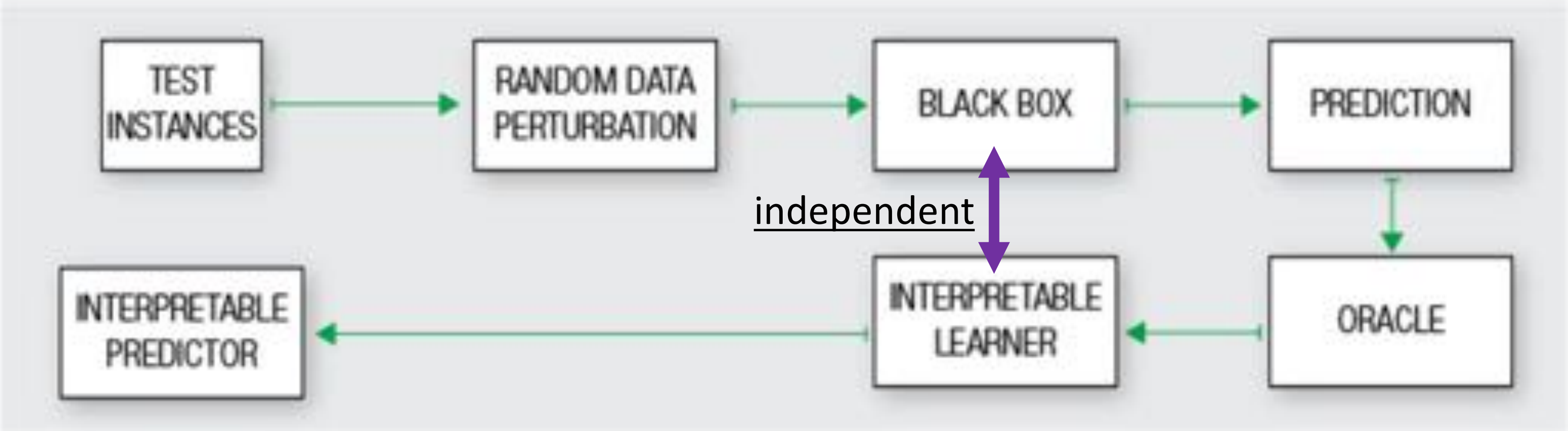
- Decision Tree (***DT***)
- Decision Rules (***DR***)
- Features Importance (***FI***)
- Saliency Mask (***SM***)
- Sensitivity Analysis (***SA***)
- Partial Dependence Plot (***PDP***)
- Prototype Selection (***PS***)
- Activation Maximization (***AM***)

Reverse Engineering

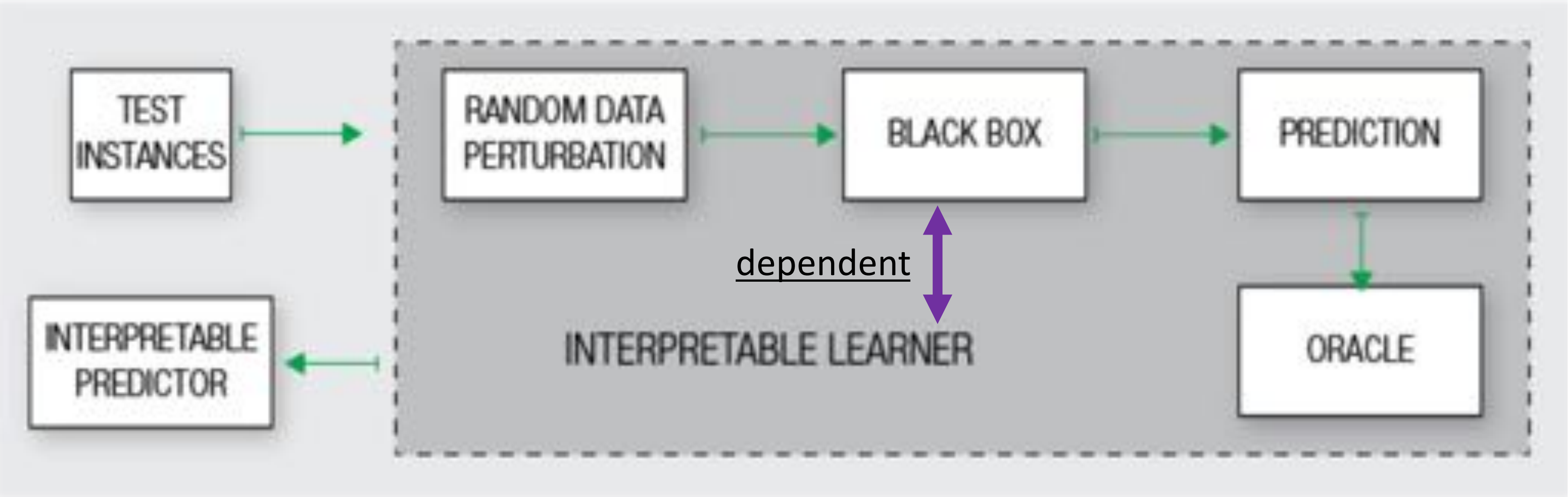
- The name comes from the fact that we can only ***observe*** the ***input*** and ***output*** of the black box.
- Possible actions are:
  - ***choice*** of a particular comprehensible predictor
  - querying/auditing the black box with input records created in a controlled way using ***random perturbations***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



Model-Agnostic vs Model-Specific

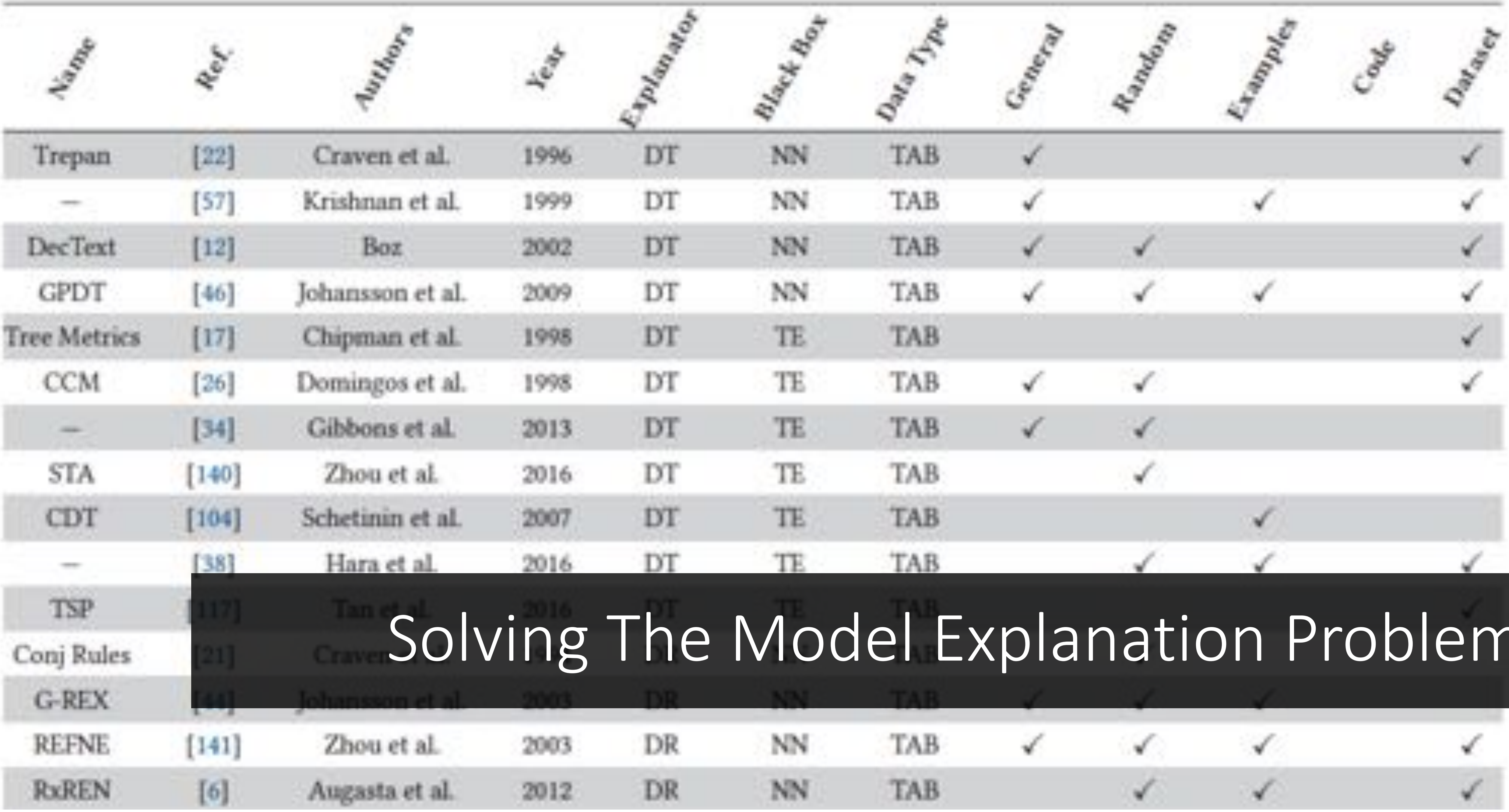


independent



dependent

Solving The Model Explanation Problem



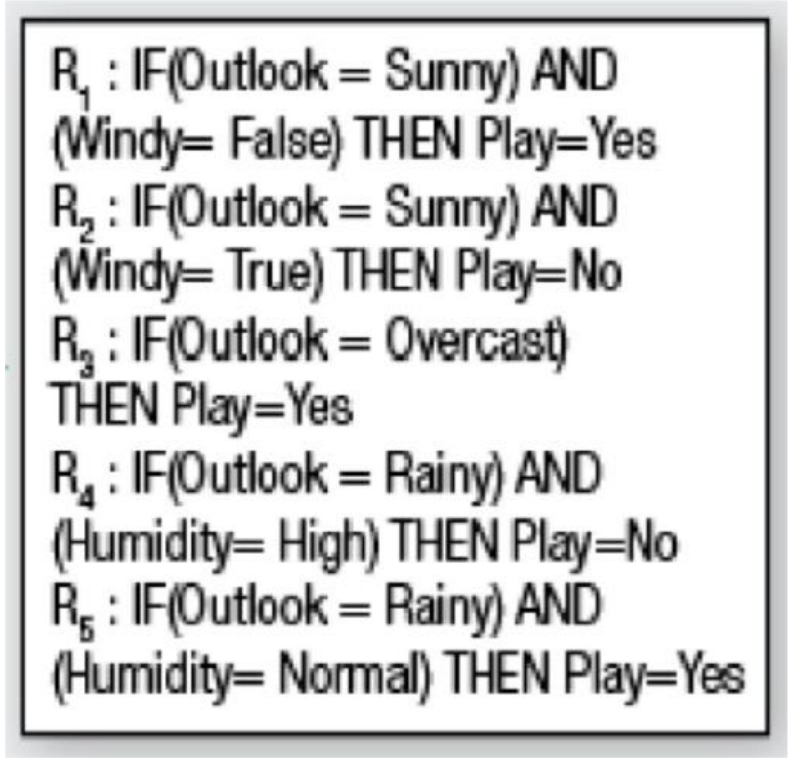
Name	Ref.	Authors	Year	Explanator	Black Box	Data Type	General	Random	Examples	Code	Dataset
Trepan	[22]	Craven et al.	1996	DT	NN	TAB	✓				✓
–	[57]	Krishnan et al.	1999	DT	NN	TAB	✓		✓		✓
DecText	[12]	Boz	2002	DT	NN	TAB	✓	✓			✓
GPDT	[46]	Johansson et al.	2009	DT	NN	TAB	✓	✓	✓		✓
Tree Metrics	[17]	Chipman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

Global Model Explainers

- Explanator: DT
    - Black Box: NN, TE
    - Data Type: TAB

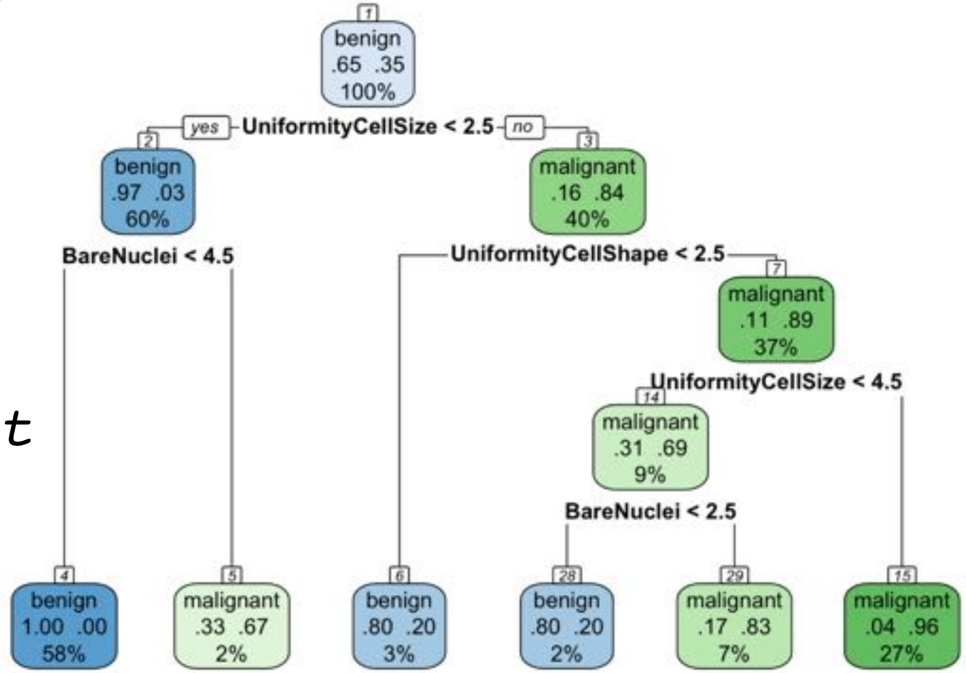
- Explanator: DR
    - Black Box: NN, SVM, TE
    - Data Type: TAB

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



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

Trepan – DT, NN, TAB

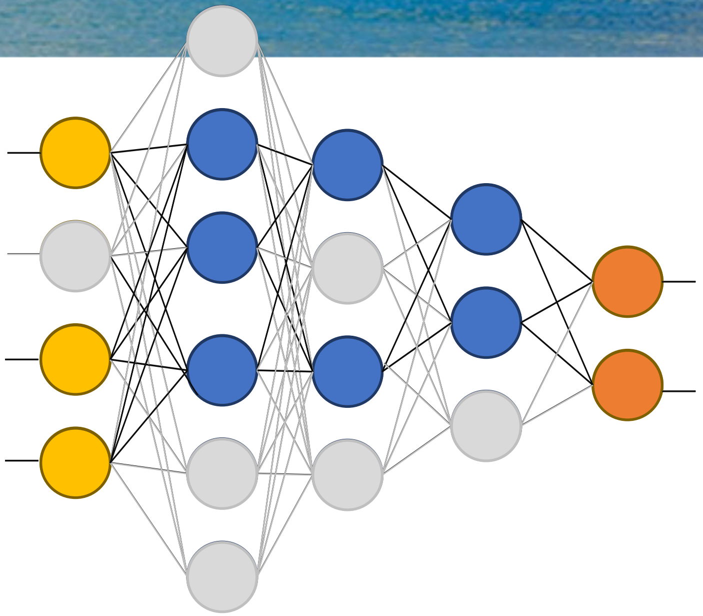


```
T = root_of_the_tree()
Q = <T, X, {}>
while Q not empty & size(T) < limit
    N, $X_N$, $C_N$ = pop(Q)
    $Z_N$ = random($X_N$, $C_N$)
    $y_Z$ = b(Z), y = b($X_N$)
    if same_class(y ∪ $y_Z$)
        continue
    S = best_split($X_N$ ∪ $Z_N$, y ∪ $y_Z$)
    S'= best_m-of-n_split(S)
    N = update_with_split(N,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

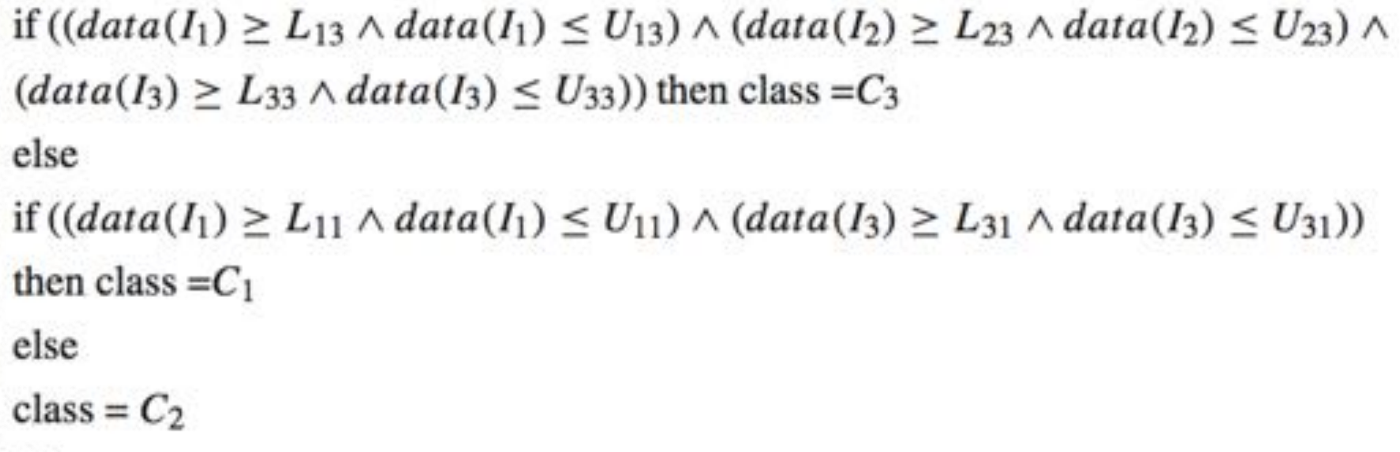
black box auditing →

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

RxREN – DR, NN, TAB



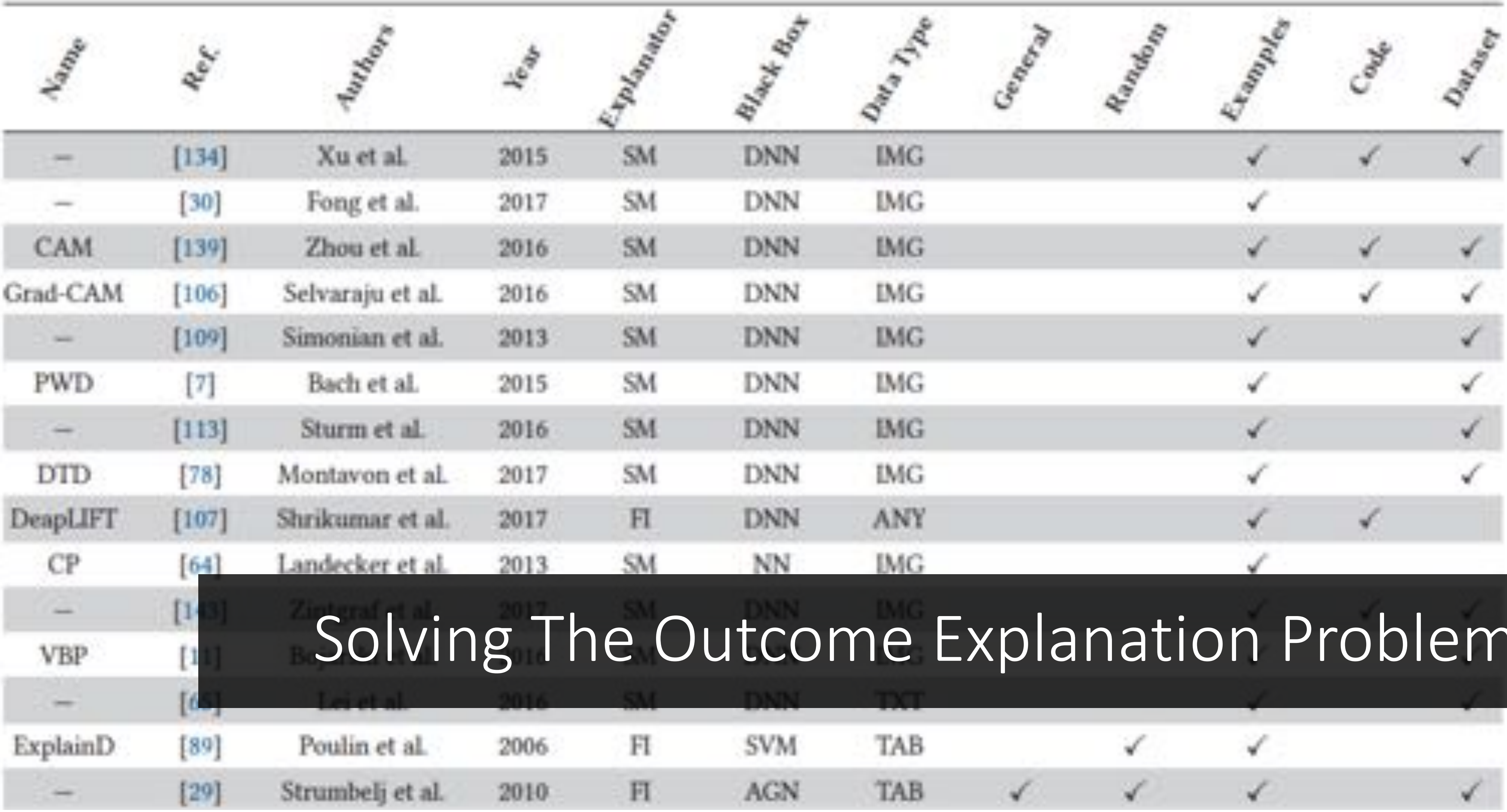
```
prune insignificant neurons
for each significant neuron
  for each outcome
    black box auditing → compute mandatory data ranges
for each outcome
  build rules using data ranges of each neuron
prune insignificant rules
update data ranges in rule conditions analyzing error
```



if $((data(I_1) \geq L_{13} \wedge data(I_1) \leq U_{13}) \wedge (data(I_2) \geq L_{23} \wedge data(I_2) \leq U_{23}) \wedge$
$(data(I_3) \geq L_{33} \wedge data(I_3) \leq U_{33}))$ then class $=C_3$
else
if $((data(I_1) \geq L_{11} \wedge data(I_1) \leq U_{11}) \wedge (data(I_3) \geq L_{31} \wedge data(I_3) \leq U_{31}))$
then class $=C_1$
else
class $= C_2$

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

Solving The Outcome Explanation Problem



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

Local Model Explainers

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

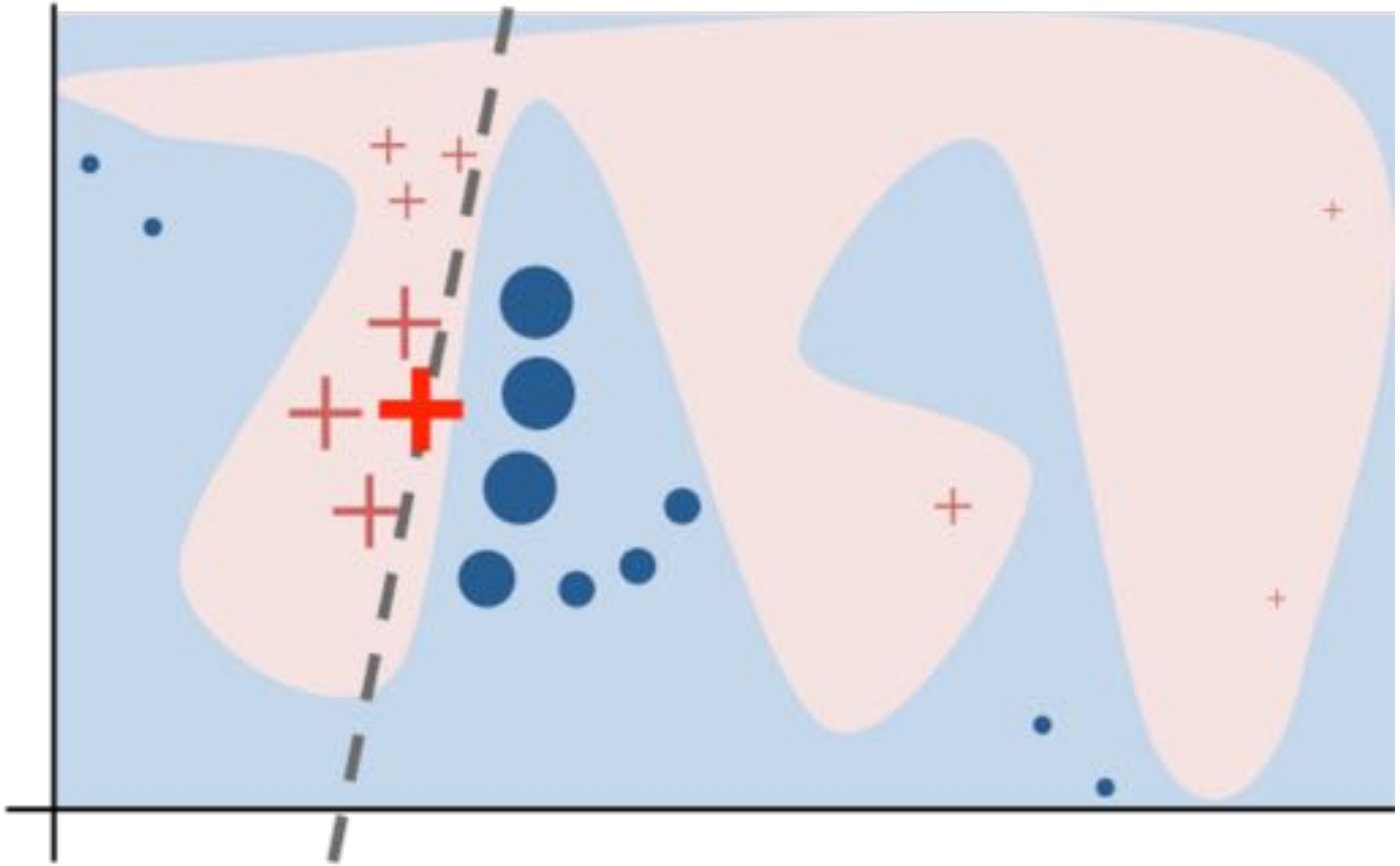
- Explanator: FI
    - Black Box: DNN, SVM
    - Data Type: ANY

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

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

Local Explanation

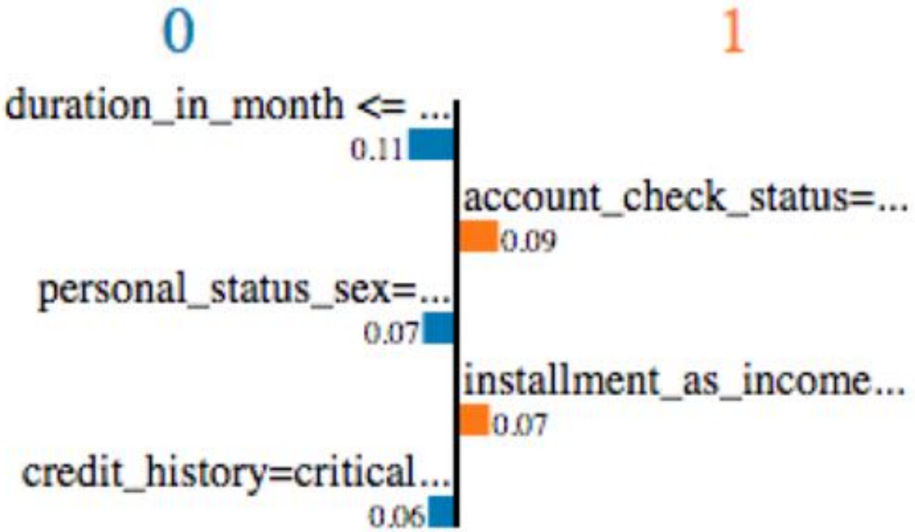
- The overall decision boundary is complex
- In the neighborhood of a single decision, the boundary is simple
- A single decision can be explained by auditing the black box around the given instance and learning a ***local*** decision.

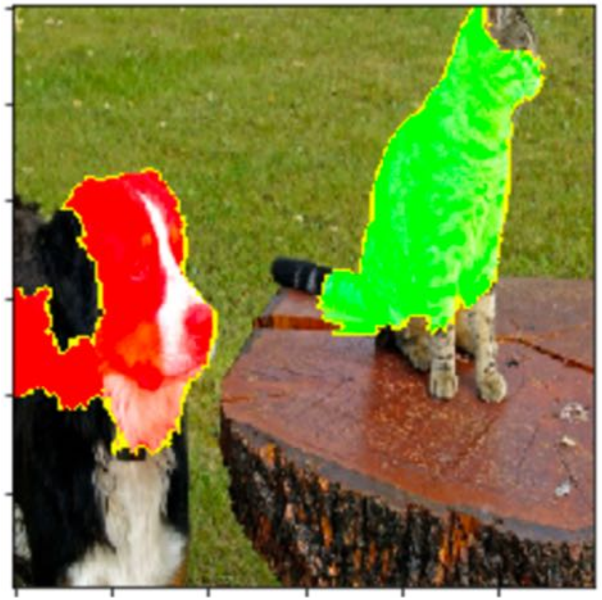


LIME – FI, AGN, ANY

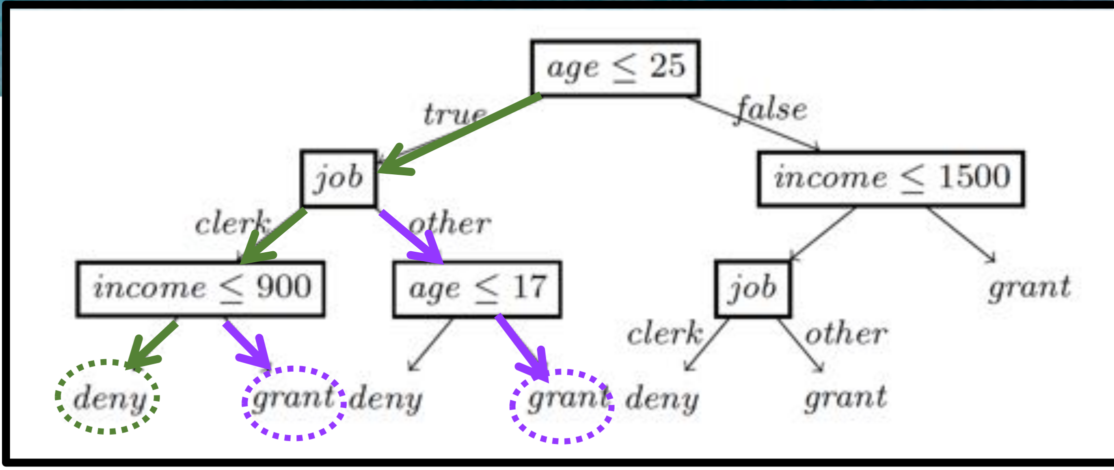
```
Z = {}
x instance to explain
x' = real2interpretable(x)
for i in {1, 2, …, N}
    $z_i$= sample_around(x')
    z = interpretabel2real(z')
    Z = Z U {<$z_i$, b($z_i$), d(x, z)>}
w = solve_Lasso(Z, k)
return w
```

black box auditing





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



LORE – DR, AGN, TAB

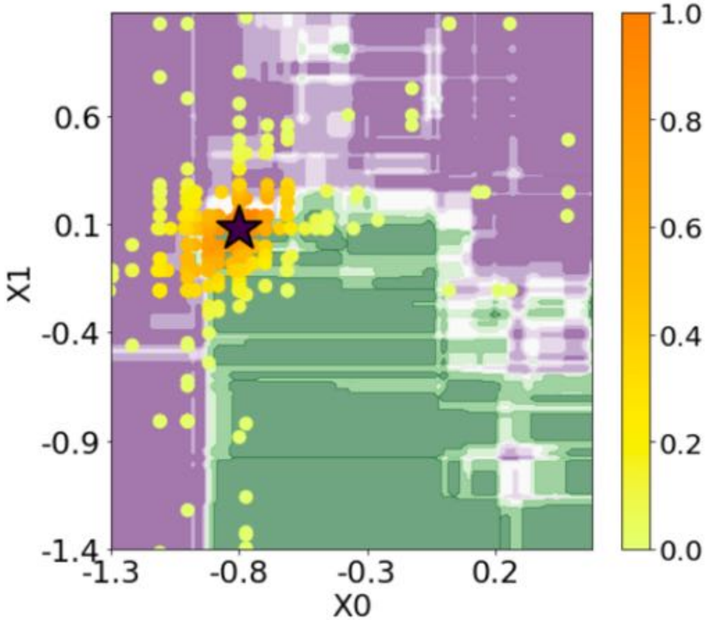
```
x instance to explain
$Z_=$ = geneticNeighborhood(x, $fitness_=$, N/2)
$Z_{\neq}$ = geneticNeighborhood(x, $fitness_{\neq}$, N/2)
Z = $Z_=$ U $Z_{\neq}$
c = buildTree(Z, b(Z))
r = (p -> y) = extractRule(c, x)
ϕ = extractCounterfactual(c, r, x)
return e = <r, ϕ>
```

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



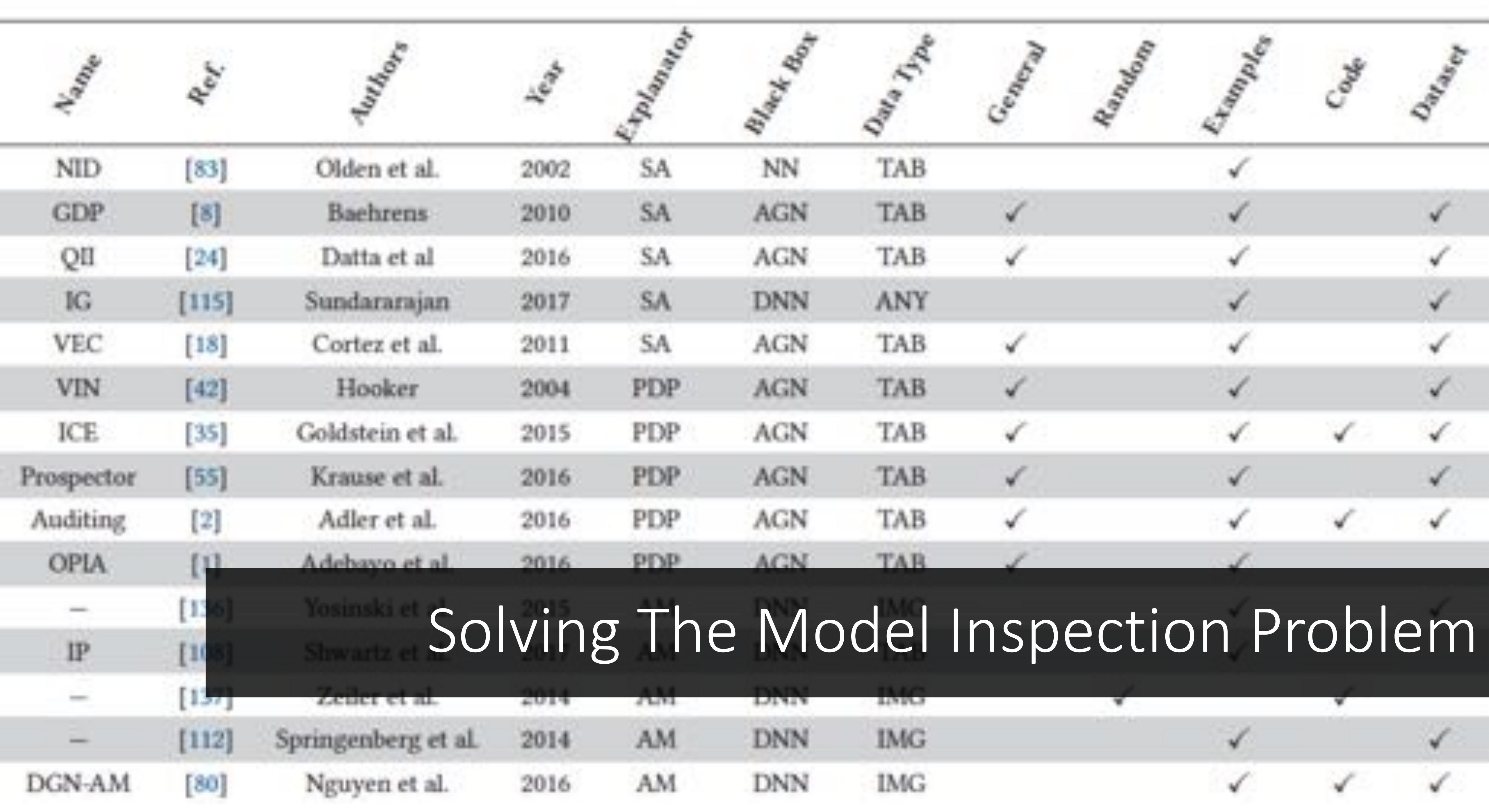
Meaningful Perturbations – SM, DNN, IMG

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

black box auditing

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

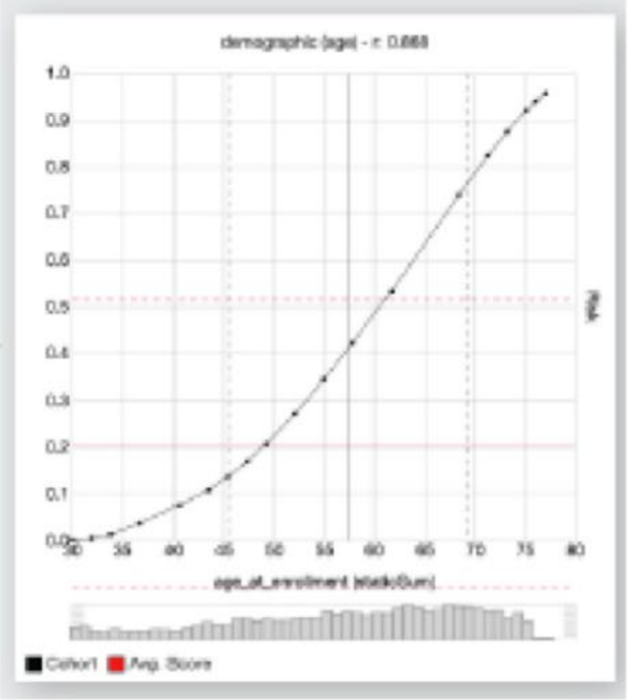
Solving The Model Inspection Problem



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

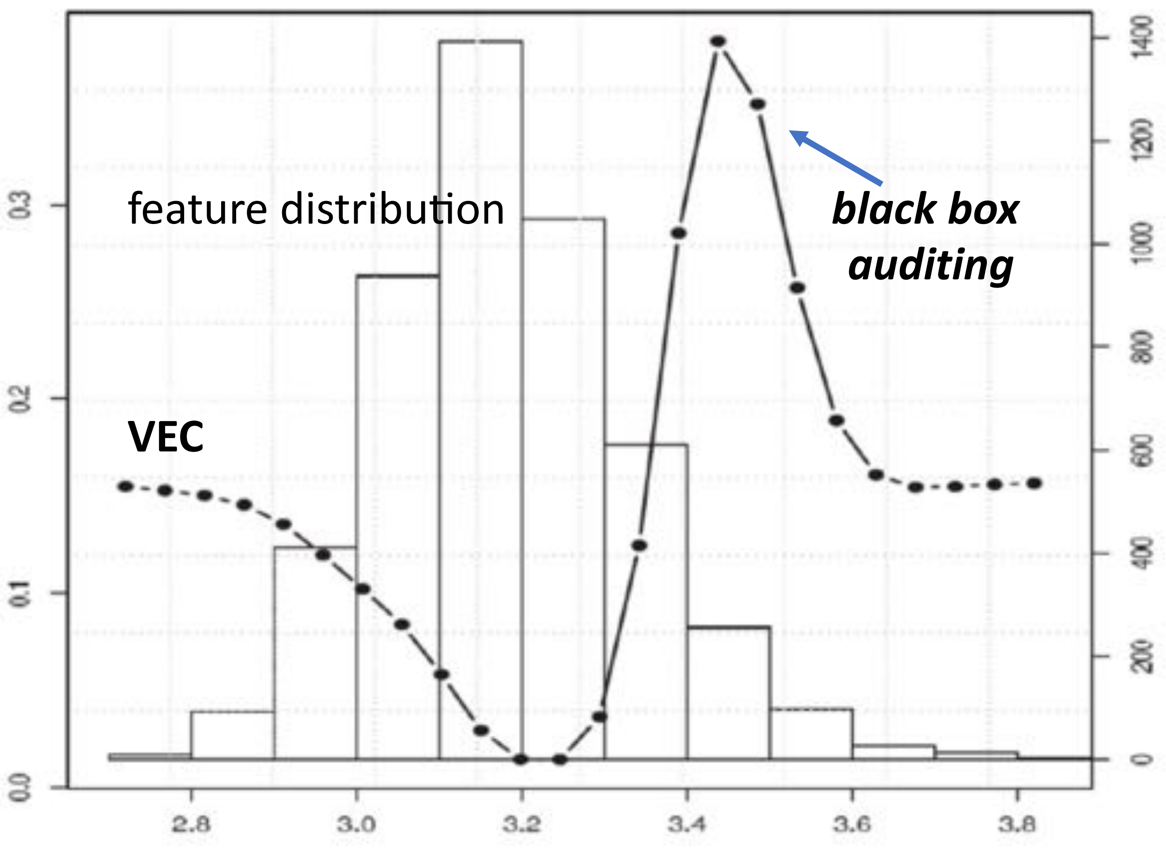
Inspection Model Explainers

- 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.


feature distribution

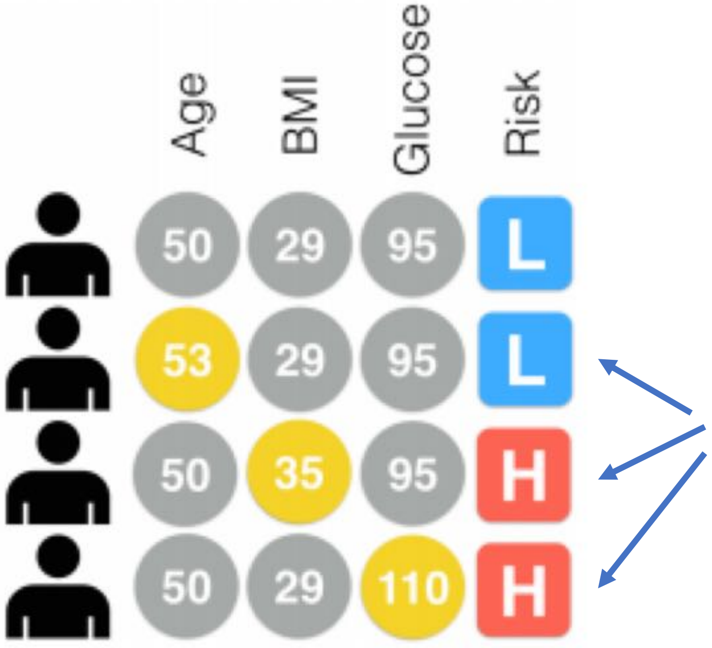
VEC

black box auditing

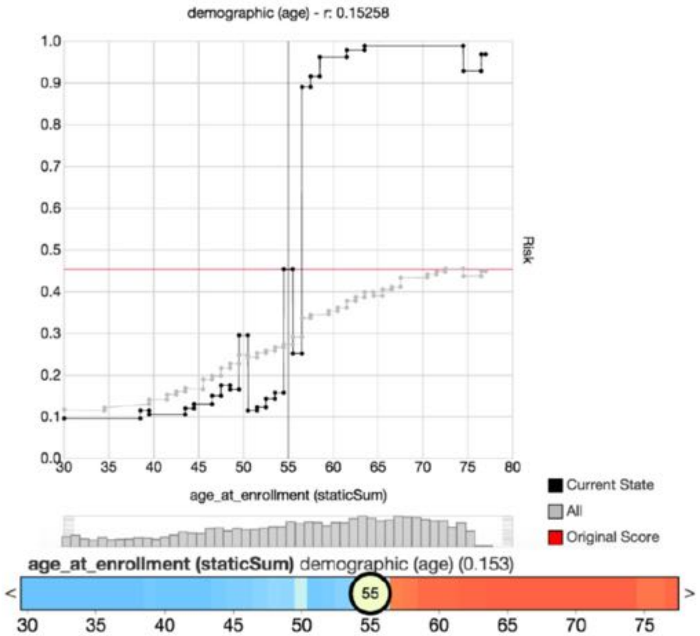
- Paulo Cortez and Mark J. Embrechts. 2011. ***Opening black box data mining models using sensitivity analysis***. CIDM.

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***.



black box auditing



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

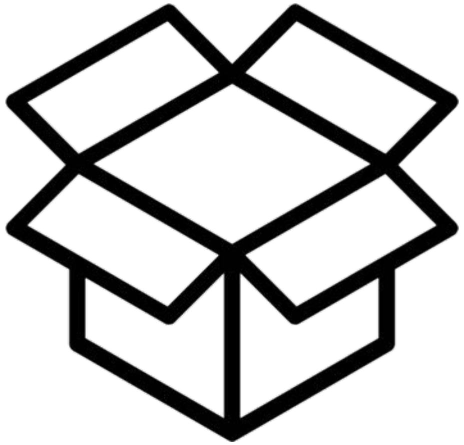
Solving The Transparent Design Problem

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			✓	✓	✓

Transparent Model Explainers

- Explanators:
  - DR
  - DT
  - PS

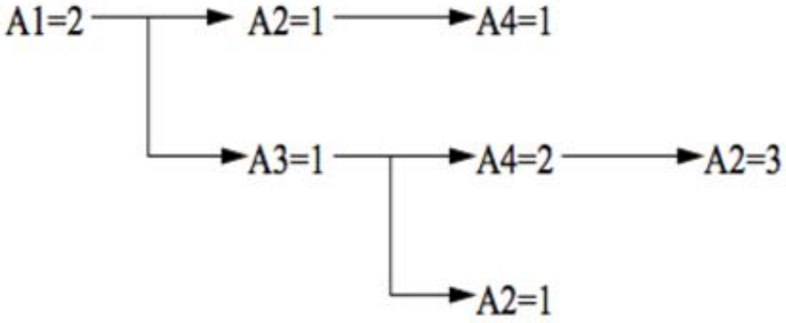
- Data Type:
  - TAB



CPAR – DR, TAB

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

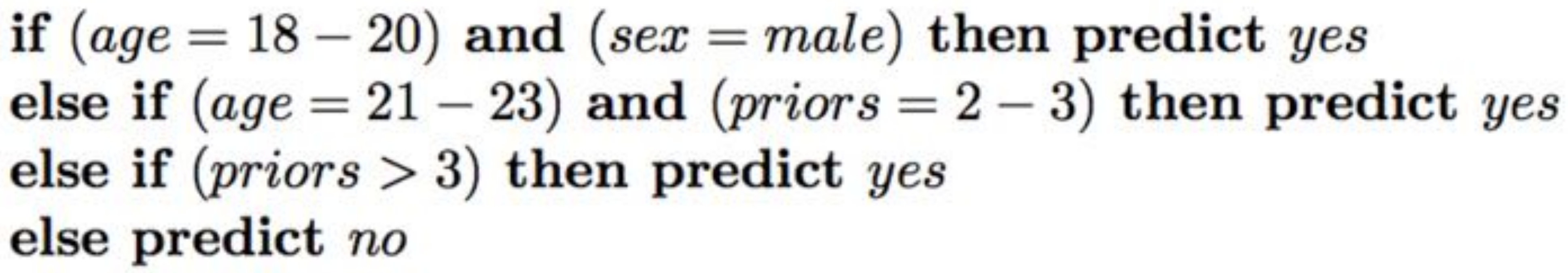
$$(A_1 = 2,\ A_2 = 1,\ A_4 = 1).$$
$$(A_1 = 2,\ A_3 = 1,\ A_4 = 2,\ A_2 = 3).$$
$$(A_1 = 2,\ A_3 = 1,\ A_2 = 1).$$



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

CORELS – DR, TAB

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



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

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

Take Home Message

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?

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.

Thank you

riccardo.guidotti@isti.cnr.it

dino.pedreschi@di.unipi.it

AI4EU

http://ai4eu.org/

SoBigData

http://www.sobigdata.eu/

HUMANE AI

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

Explanation with Background Information

Md Kamruzzaman Sarker

Pascal Hitzler

Wright State University

DaSe Lab

SSC

Ohio Federal Research Network (OFRN)

Explanation with Background Knowledge

- We tend to give explanation in terms of our current knowledge.
- From our childhood we learn that dog has 4 legs, 1 head, 1 tongue, 1 tail etc.
- When we see any image of dog our thinking automatically try to capture those objects.
- We always want to conform with our previously acquired knowledge (Background Knowledge).

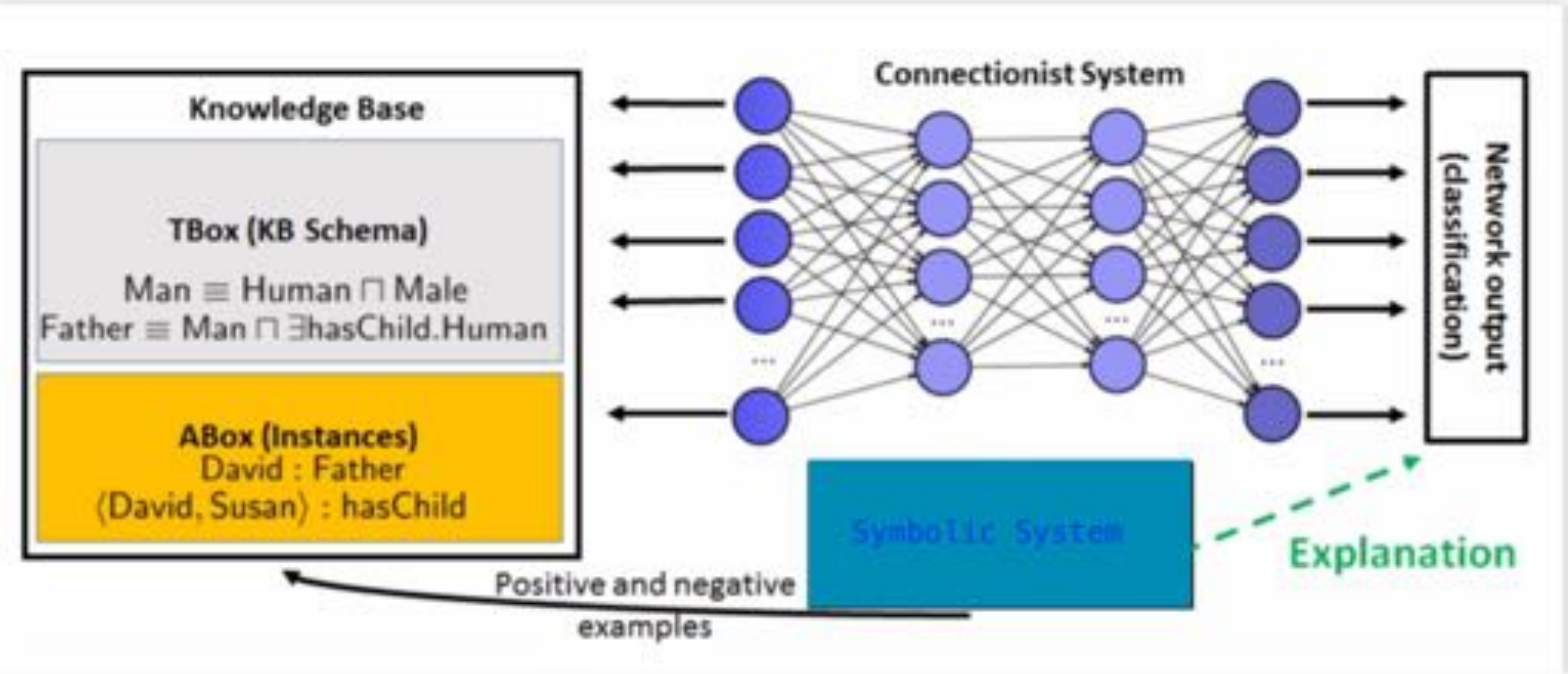
Will not it be better if we can explain in terms of our knowledge?

How ?

Hard to make connection between our knowledge and a model which is trained by reducing loss.

Idea found in current literature is similar to inductive programming.

- Use background knowledge in the form of linked data and ontologies to help explain.
- Link inputs and outputs to background knowledge.
- Use a symbolic learning system to generate an explanatory theory.


Knowledge Base
TBox (KB Schema)
$\text{Man} \equiv \text{Human} \sqcap \text{Male}$
$\text{Father} \equiv \text{Man} \sqcap \exists \text{hasChild.Human}$
ABox (Instances)
David : Father
$\langle \text{David, Susan} \rangle$: hasChild
Connectionist System
Network output (classification)
Symbolic System
Explanation
Positive and negative examples

Current symbolic systems

- ECII[1]
- DL – Learner[2]
- OWL Miner[3]
- DL – Miner[4]

Input Needed for These Systems

- Background information/Ontology/Knowledge Graphs
- Some positive and/or negative examples
- Mapping between model dataset and the ontology

Real-world Background Info as Knowledge Graphs

- Cyc
- Wordnet
- Suggested Merged Upper Ontology (SUMO)
- Dbpedia
- Freebase

Positive & Negative

- The concept is considered is positive and all others are negative.[5]

Mapping between dataset and Ontology

- Mapping each instance as an individual and put it in exact hierarchy.[5]

Experiment using MIT ADE20K-Dataset

http://groups.csail.mit.edu/vision/datasets/ADE20K/

Experiment using MIT ADE20K-Dataset

Images come with annotations of objects in the picture:

```
001 # 0 # 0 # sky # sky # ""
002 # 0 # 0 # road, route # road # ""
005 # 0 # 0 # sidewalk, pavement # sidewalk # ""
006 # 0 # 0 # building, edifice # building # ""
007 # 0 # 0 # truck, motortruck # truck # ""
008 # 0 # 0 # hovel, hut, shack, shanty # hut # "’
009 # 0 # 0 # pallet # pallet # ""
011 # 0 # 0 # box # boxes # ""
001 # 1 # 0 # door # door # ""
002 # 1 # 0 # window # window # ""
009 # 1 # 0 # wheel # wheel # ""
```

Mapping

Objects in image annotations became individuals (constants), which can be typed with the ontology.

```
contains road1
contains window1
contains door1
contains wheel1
contains sidewalk1
contains truck1
contains box1
contains building1
```

Proof of Concept Experiment

Positive Examples (Outdoor Warehouse)

Negative Examples (Indoor Warehouse)

Proof of Concept Experiment

Positive:

img1: road, window, door, wheel, sidewalk, truck, box, building
img2: tree, road, window, timber, building, lumber
img3: hand, sidewalk, clock, steps, door, face, building, window, road

Negative:

img4: shelf, ceiling, floor

img5: box, floor, wall, ceiling, product

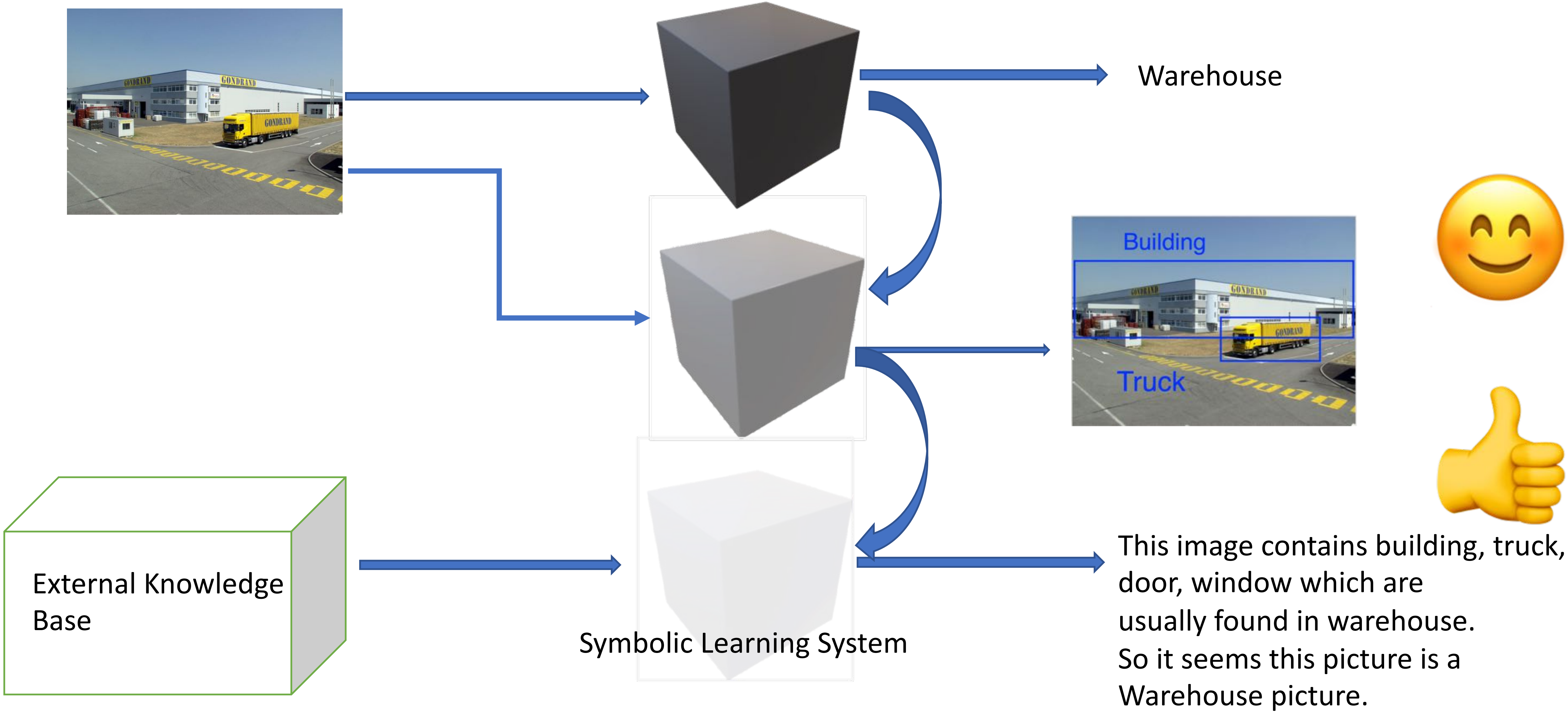
img6: ceiling, wall, shelf, floor, product

results include:

$$\exists \mathrm{contains}.\mathrm{Transitway}$$

$$\exists \mathrm{contains}.\mathrm{LandArea}$$

DL Model which merges explanation with Background information



Warehouse

External Knowledge Base

Symbolic Learning System

This image contains building, truck, door, window which are usually found in warehouse. So it seems this picture is a Warehouse picture.

Summary

- This is just beginning of using background information to enhance explanation.
- There are many open questions-
  - Where we can get effective background information?
  - How to relate already available background information with my current model?
  - Are those explanations enough to satisfy our quest?

References

1. Md Kamruzzaman Sarker, Pascal Hitzler, 2019. Concept Induction for description logics, AAAI-19.

2. Jens Lehmann, and Pascal Hitzler, 2010. Concept learning in de- scription logics using refinement operators. *Machine Learn- ing* 78(1-2):203–250

3. David Ratcliffe and Kerry Taylor, 2016. Closed-World Concept Induction for Learning in OWL Knowledge Bases. EKAW - 2016

4. Viachaslau Sazonau, 2017. General Terminology Induction in Description Logics, PhD Thesis.

5. Md Kamruzzaman Sarker, Pascal Hitzler, 2017. Explaining Input Output Relationship of Training Neural Networks : First Steps, Nesy 2017.

Machine Learning in Knowledge Graphs

Pasquale Minervini
University College London / UCL NLP
@pminervini

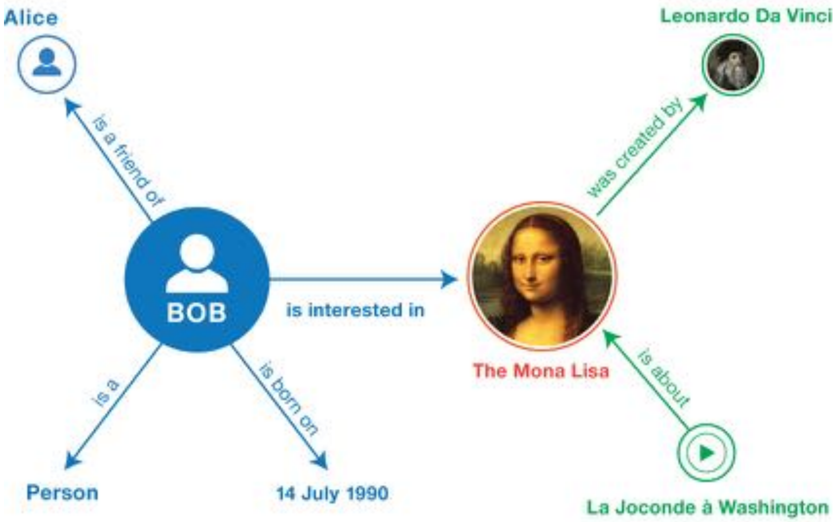
Outline

- Knowledge Graphs
    - What are they?
    - Applications in Industry and Academia
    - Problems with building large-scale Knowledge Graphs
- Relational Learning in Knowledge Graphs
    - Observable Feature Models
    - Latent Feature Models
    - Combining and Interpreting Observable and Latent Feature Models
- Neuro-Symbolic Reasoning

Knowledge Graphs

- Set of (*subject*, *predicate*, *object — SPO*) **triples** - *subject* and *object* are **entities**, and *predicate* is the **relationship** holding between them.
- Each SPO **triple** denotes a **fact**, i.e. the existence of an actual relationship between two entities.

subject	**predicate**	**object**
Bob	*is interested in*	*The Mona Lisa*
Bob	*is a friend of*	*Alice*
The Mona Lisa	*was created by*	*Leonardo Da Vinci*
Bob	*is a*	*Person*
La Joconde à W.	*is about*	*The Mona Lisa*
Bob	*is born on*	*14 July 1990*



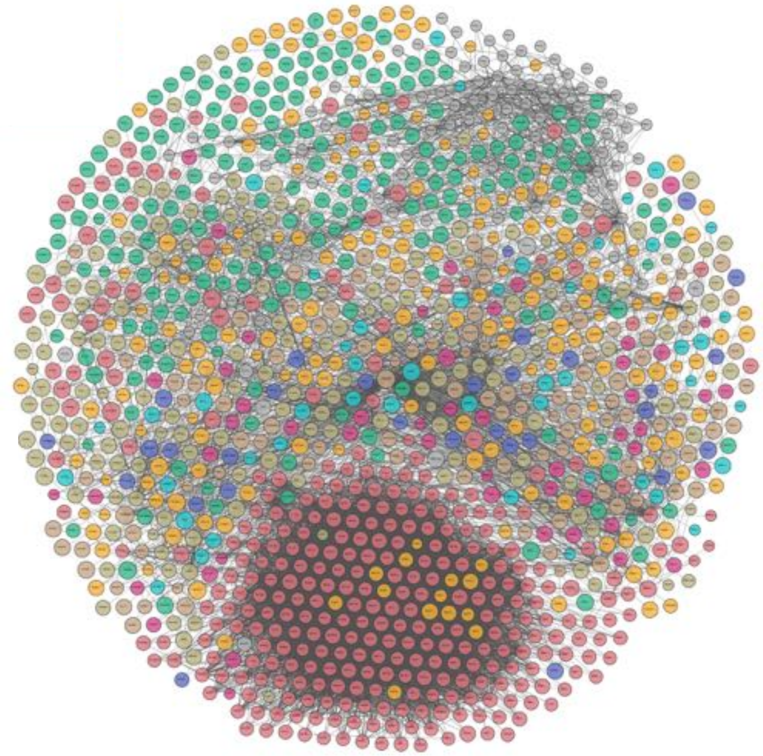
Knowledge Graphs

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
NELL	2M	425	285	433K
Google KG	570M	35K	1.5K	18B
Knowledge Vault	45M	4.5K	1.1K	271M
Yahoo! KG	3.4M	800	250	1.39B

- **Manual Construction** - curated, collaborative
- **Automated Construction** - semi-structured, unstructured

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

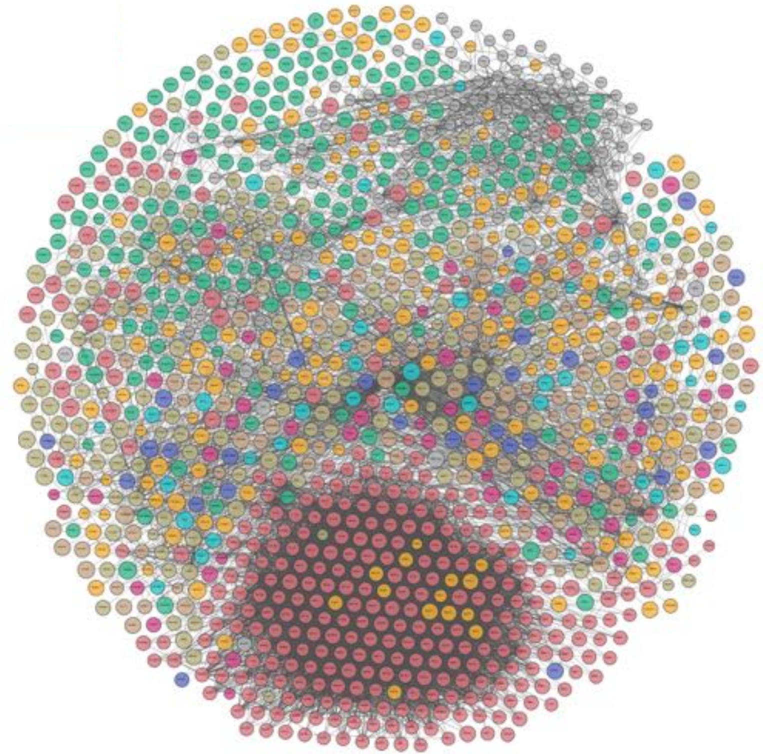
Spans a variety of domains - Geography, Government, Life Sciences, Linguistics, Media, Publications, Cross-domain..



Knowledge Graphs

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
NELL	2M	425	285	433K
Google KG	570M	35K	1.5K	18B
Knowledge Vault	45M	4.5K	1.1K	271M
Yahoo! KG	3.4M	800	250	1.39B

- **Manual Construction** - curated, collaborative
- **Automated Construction** - semi-structured, unstructured



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

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

Knowledge Graphs Construction

Knowledge Graph construction methods can be classified in:

- **Manual** — curated (e.g. via experts), collaborative (e.g. via volunteers)
- **Automated** — semi-structured (e.g. from infoboxes), unstructured (e.g. from text)

Coverage is an issue:

- **Freebase** (40M entities) - 71% of persons without a birthplace, 75% without a nationality, even worse for other relation types [Dong et al. 2014]
- **DBpedia** (20M entities) - 61% of persons without a birthplace, 58% of scientists missing why they are popular [Krompaß et al. 2015]

Relational Learning can help us overcoming these issues.

Relational Learning in Knowledge Graphs

- **Dyadic Multi-Relational Data** [Nickel et al. 2015, Getoor et al. 2007]

- Many possible relational learning tasks:
    - **Link Prediction** — Identify missing relationships between entities
    - **Collective Classification** — Classify entities based on their relationships
    - **Link-Based Clustering** — Cluster entities based on their relationships
    - **Entity Resolution** — Entity mapping/deduplication

Relational structure is a rich source of information.

In general, the *i.i.d. assumption* does not hold in this context.

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 $P\left(\overline{\mathbf{Y}}\right)$, we end up with *three model classes*:

- **Latent Feature Models**: variables $y_{spo} \in \{0,1\}$ are *conditionally independent* given the *latent features* $\mathbf{\Theta}$ associated with subject, predicate, and object:

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

- **Observable Feature Models**: related to Latent Feature Models, but $\mathbf{\Theta}$ 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* $f\left(s, p, o \mid \Theta\right)$ representing the likelihood that a triple is in the KG, conditioned on the parameters $\boldsymbol{\Theta}$:

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

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

Observable Feature Models - Uni-Relational Similarities

Uni-Relational Similarity Measures: based on *homophily* — similar entities are likely to be related — and *neighbourhood similarity.*

- **Local**: derive similarity between entities from their local neighbourhood
  (e.g. Common Neighbours, Adamic-Adar Index [Adamic et al. 2003], Preferential Attachment [Barabási et al. 1999], ..)

- **Global**: derive similarity between entities using the whole graph
  (e.g. Katz Index [Katz, 1953], Leicht-Holme-Newman Index [Leicht et al. 2006], PageRank [Brin et al. 1998], ..)

- **Quasi-Local**: trade-off between computational complexity and predictive accuracy
  (e.g. Local Katz Index [Liben-Nowell et al. 2007], Local Random Walks [Liu et al. 2010], ..)

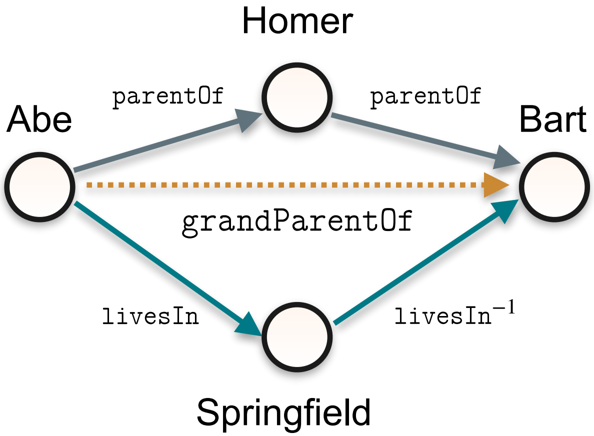
Observable Feature Models - Rule Mining and ILP

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:

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

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]

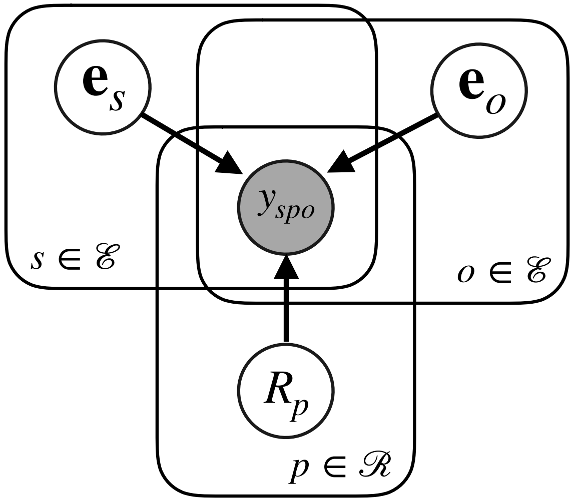
Observable Feature Models are *Interpretable*

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

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

Latent Feature Models

Variables y_{spo} are conditionally independent given a set of latent features and parameters $\mathbf{\Theta}$. *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 <u>very hard to interpret</u>.

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]	$- \left\lVert \mathbf{e}_s + \mathbf{r}_p - \mathbf{e}_o \right\rVert_{1,2}^{2}$	$\mathbf{r}_p \in \mathbb{R}^k$
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 [Nickel 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]	$\boxed{\mathbf{e}_s^\top \mathbf{W}_p \mathbf{e}_o}$	$\mathbf{W}_p \in \mathbb{R}^{k\times k}$
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]	$-\left\lVert \mathbf{e}_s + \mathbf{r}_p - \mathbf{e}_o \right\rVert_{1,2}^{2}$	$\mathbf{r}_p \in \mathbb{R}^k$
DistMult [Yang et al. 2015]	$\langle \mathbf{e}_s, \mathbf{r}_p, \mathbf{e}_o \rangle$	$\mathbf{r}_p \in \mathbb{R}^k$
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 [Nickel et al. 2016]	$\text{Re}\left(\langle \mathbf{e}_s, \mathbf{r}_p, \bar{\mathbf{e}}_o \rangle\right)$	$\mathbf{r}_p \in \mathbb{C}^k$
ConvE [Dettmers et al. 2017]	$f\left(\text{vec}\left(f\left([\overline{\mathbf{e}_s}; \overline{\mathbf{r}_p}] * \omega\right)\right) \mathbf{W}\right) \mathbf{e}_o$	$\mathbf{r}_p \in \mathbb{R}^k, \mathbf{W} \in \mathbb{R}^{c\times k}$

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]	$- \parallel \mathbf{e}_s + \mathbf{r}_p - \mathbf{e}_o \parallel_{1,2}^{2}$	$\mathbf{r}_p \in \mathbb{R}^k$
DistMult [Yang et al. 2015]	$\langle \mathbf{e}_s, \mathbf{r}_p, \mathbf{e}_o \rangle$	$\mathbf{r}_p \in \mathbb{R}^k$
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 [Nickel et al. 2016]	$\text{Re} \left(\langle \mathbf{e}_s, \mathbf{r}_p, \bar{\mathbf{e}}_o \rangle \right)$	$\mathbf{r}_p \in \mathbb{C}^k$
ConvE [Dettmers et al. 2017]	$f \left(\text{vec} \left(f \left([\overline{\mathbf{e}_s}; \overline{\mathbf{r}_p}] * \omega \right) \right) \mathbf{W} \right) \mathbf{e}_o$	$\mathbf{r}_p \in \mathbb{R}^k, \mathbf{W} \in \mathbb{R}^{c \times k}$

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]	$\boxed{\mathbf{e}_s^\top \mathbf{W}_p \mathbf{e}_o}$	$\mathbf{W}_p \in \mathbb{R}^{k \times k}$
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]	$- \left\Vert \mathbf{e}_s + \mathbf{r}_p - \mathbf{e}_o \right\Vert_{1,2}^{2}$	$\mathbf{r}_p \in \mathbb{R}^k$
DistMult [Yang et al. 2015]	$\boxed{\langle \mathbf{e}_s, \mathbf{r}_p, \mathbf{e}_o \rangle}$	$\mathbf{r}_p \in \mathbb{R}^k$
HolE [Nickel et al. 2016]	$\mathbf{r}_p^\top \left(\mathscr{F}^{-1}\left[\overline{\mathscr{F}[\mathbf{e}_s]} \odot \mathscr{F}[\mathbf{e}_o]\right]\right)$	$\mathbf{r}_p \in \mathbb{R}^k$
ComplEx [Nickel et al. 2016]	$\boxed{\text{Re}\left(\langle \mathbf{e}_s, \mathbf{r}_p, \bar{\mathbf{e}}_o \rangle\right)}$	$\mathbf{r}_p \in \mathbb{C}^k$
ConvE [Dettmers et al. 2017]	$f\left(\text{vec}\left(f\left([\overline{\mathbf{e}_s}; \overline{\mathbf{r}_p}] * \omega\right)\right)\mathbf{W}\right)\mathbf{e}_o$	$\mathbf{r}_p \in \mathbb{R}^k, \mathbf{W} \in \mathbb{R}^{c \times k}$

Latent Feature Models - Learning

Another core differente among models is the *loss function* minimised for fitting the latent parameters $\boldsymbol{\Theta}$ to the data — let $f_{spo} = f\left(x_{spo} \mid \Theta\right)$ and $p_{spo} = \sigma\left(f_{spo}\right)$:

Losses	**Formulation**	**Models**
Quadratic Loss	$\sum_{(x_{spo},y_{spo})\in\mathcal{D}} \left\| y_{spo} - f_{spo} \right\|_2^2$	Tensor Factorisation, RESCAL (ALS)
Pairwise Loss	$\sum_{x_+\in\mathcal{D}_+}\sum_{x_-\in\mathcal{D}_-} \mathcal{L}(x_+, x_-) \overset{e.g.}{=} \max\left\{0,\gamma + f_{x_-} - f_{x_+}\right\}$	SE, NTN, TransE, HolE
Cross-Entropy Loss	$\sum_{(x,y)\in\mathcal{D}} \left[y \log\left(p_x\right) + (1-y)\log\left(1-p_x\right)\right]$	ComplEx
Multiclass Loss	$\sum_{x_{spo}\in\mathcal{D}_+} \mathcal{L}(p_{spo},1) + \sum_{\tilde{s}\in\mathcal{E}} \mathcal{L}(p_{\tilde{s}po}, y_{\tilde{s}po}) + \sum_{\tilde{o}\in\mathcal{E}} \mathcal{L}(p_{sp\tilde{o}}, y_{sp\tilde{o}})$	ConvE, ComplEx-N3 [Dettmers et al. 2017, Lacroix et al. 2018]

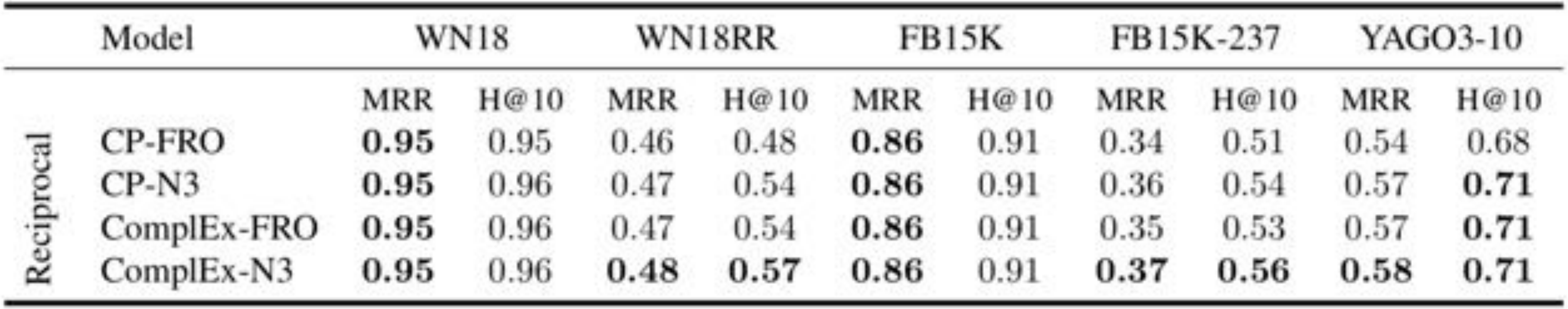
Latent Feature Models - Predictive Accuracy

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.

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

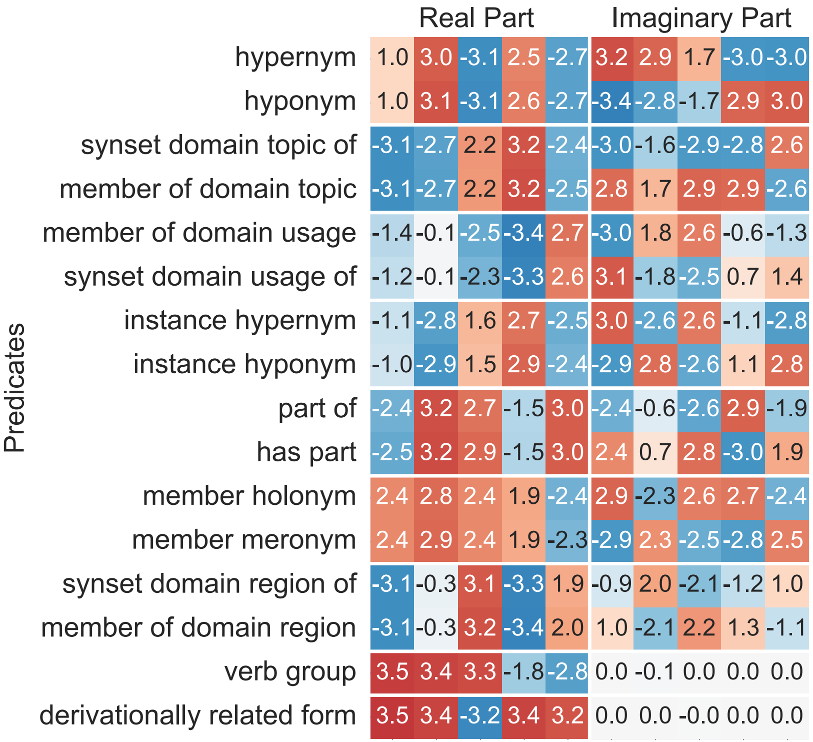
From [Lacroix et al. ICML 2018]



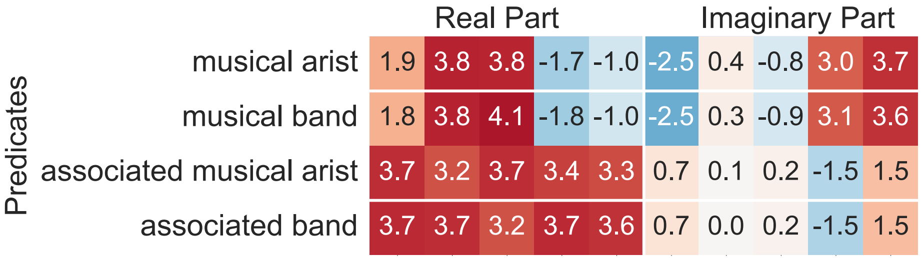
	Model	WN18		WN18RR		FB15K		FB15K-237		YAGO3-10	
		MRR	H@10	MRR	H@10	MRR	H@10	MRR	H@10	MRR	H@10
Reciprocal	CP-FRO	**0.95**	0.95	0.46	0.48	**0.86**	0.91	0.34	0.51	0.54	0.68
	CP-N3	**0.95**	0.96	0.47	0.54	**0.86**	0.91	0.36	0.54	0.57	**0.71**
	ComplEx-FRO	**0.95**	0.96	0.47	0.54	**0.86**	0.91	0.35	0.53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

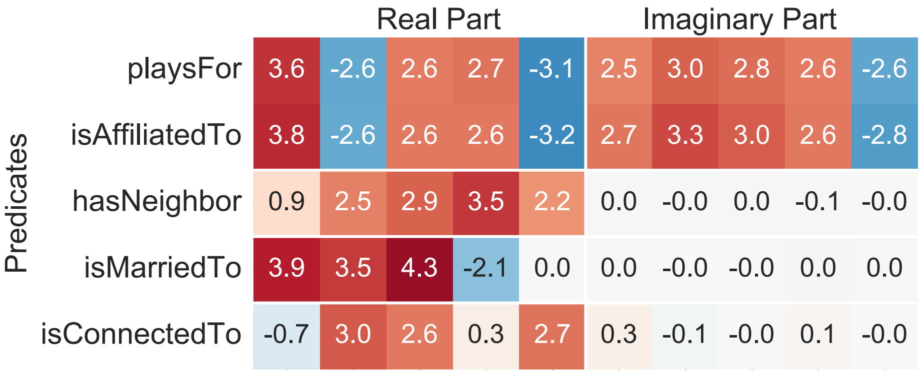
Learned relation embeddings — using *ComplEx* with a *pairwise margin-based loss* — for WordNet (left), DBpedia, and YAGO (right) [Minervini et al. ECML 2017]



Predicates	Real Part					Imaginary Part				
hypernym	1.0	3.0	-3.1	2.5	-2.7	3.2	2.9	1.7	-3.0	-3.0
hyponym	1.0	3.1	-3.1	2.6	-2.7	-3.4	-2.8	-1.7	2.9	3.0
synset domain topic of	-3.1	-2.7	2.2	3.2	-2.4	-3.0	-1.6	-2.9	-2.8	2.6
member of domain topic	-3.1	-2.7	2.2	3.2	-2.5	2.8	1.7	2.9	2.9	-2.6
member of domain usage	-1.4	-0.1	-2.5	-3.4	2.7	-3.0	1.8	2.6	-0.6	-1.3
synset domain usage of	-1.2	-0.1	-2.3	-3.3	2.6	3.1	-1.8	-2.5	0.7	1.4
instance hypernym	-1.1	-2.8	1.6	2.7	-2.5	3.0	-2.6	2.6	-1.1	-2.8
instance hyponym	-1.0	-2.9	1.5	2.9	-2.4	-2.9	2.8	-2.6	1.1	2.8
part of	-2.4	3.2	2.7	-1.5	3.0	-2.4	-0.6	-2.6	2.9	-1.9
has part	-2.5	3.2	2.9	-1.5	3.0	2.4	0.7	2.8	-3.0	1.9
member holonym	2.4	2.8	2.4	1.9	-2.4	2.9	-2.3	2.6	2.7	-2.4
member meronym	2.4	2.9	2.4	1.9	-2.3	-2.9	2.3	-2.5	-2.8	2.5
synset domain region of	-3.1	-0.3	3.1	-3.3	1.9	-0.9	2.0	-2.1	-1.2	1.0
member of domain region	-3.1	-0.3	3.2	-3.4	2.0	1.0	-2.1	2.2	1.3	-1.1
verb group	3.5	3.4	3.3	-1.8	-2.8	0.0	-0.1	0.0	0.0	0.0
derivationally related form	3.5	3.4	-3.2	3.4	3.2	0.0	0.0	-0.0	0.0	0.0



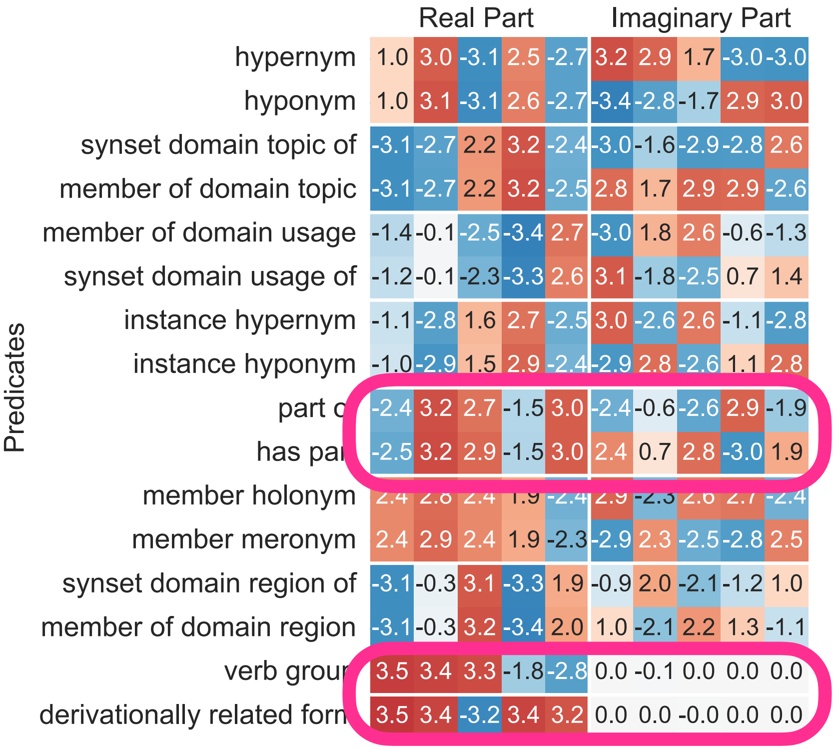
Predicates	Real Part					Imaginary Part				
musical arist	1.9	3.8	3.8	-1.7	-1.0	-2.5	0.4	-0.8	3.0	3.7
musical band	1.8	3.8	4.1	-1.8	-1.0	-2.5	0.3	-0.9	3.1	3.6
associated musical arist	3.7	3.2	3.7	3.4	3.3	0.7	0.1	0.2	-1.5	1.5
associated band	3.7	3.7	3.2	3.7	3.6	0.7	0.0	0.2	-1.5	1.5



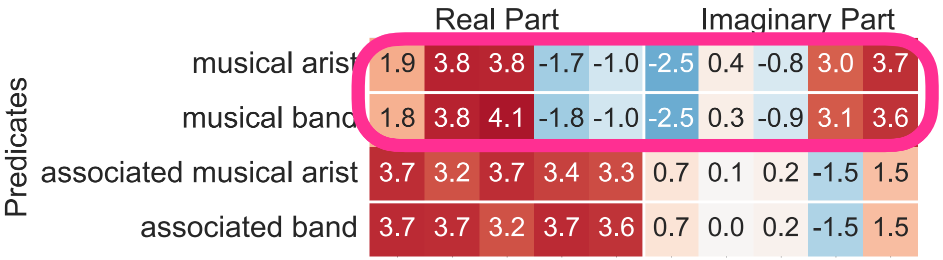
Predicates	Real Part					Imaginary Part				
playsFor	3.6	-2.6	2.6	2.7	-3.1	2.5	3.0	2.8	2.6	-2.6
isAffiliatedTo	3.8	-2.6	2.6	2.6	-3.2	2.7	3.3	3.0	2.6	-2.8
hasNeighbor	0.9	2.5	2.9	3.5	2.2	0.0	-0.0	0.0	-0.1	-0.0
isMarriedTo	3.9	3.5	4.3	-2.1	0.0	0.0	-0.0	-0.0	0.0	0.0
isConnectedTo	-0.7	3.0	2.6	0.3	2.7	0.3	-0.1	-0.0	0.1	-0.0

Latent Feature Models - Interpreting the Embeddings

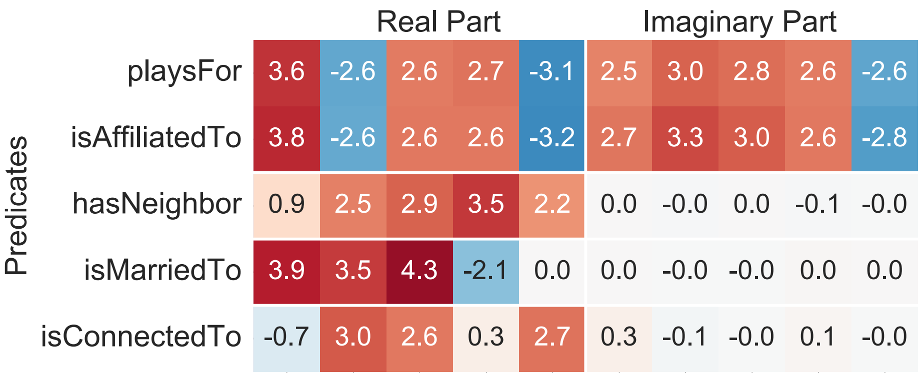
Learned relation embeddings — using *ComplEx* with a *pairwise margin-based loss* — for WordNet (left), DBpedia, and YAGO (right) [Minervini et al. ECML 2017]



Predicates	Real Part					Imaginary Part				
hypernym	1.0	3.0	-3.1	2.5	-2.7	3.2	2.9	1.7	-3.0	-3.0
hyponym	1.0	3.1	-3.1	2.6	-2.7	-3.4	-2.8	-1.7	2.9	3.0
synset domain topic of	-3.1	-2.7	2.2	3.2	-2.4	-3.0	-1.6	-2.9	-2.8	2.6
member of domain topic	-3.1	-2.7	2.2	3.2	-2.5	2.8	1.7	2.9	2.9	-2.6
member of domain usage	-1.4	-0.1	-2.5	-3.4	2.7	-3.0	1.8	2.6	-0.6	-1.3
synset domain usage of	-1.2	-0.1	-2.3	-3.3	2.6	3.1	-1.8	-2.5	0.7	1.4
instance hypernym	-1.1	-2.8	1.6	2.7	-2.5	3.0	-2.6	2.6	-1.1	-2.8
instance hyponym	-1.0	-2.9	1.5	2.9	-2.4	-2.9	2.8	-2.6	1.1	2.8
part o	-2.4	3.2	2.7	-1.5	3.0	-2.4	-0.6	-2.6	2.9	-1.9
has par	-2.5	3.2	2.9	-1.5	3.0	2.4	0.7	2.8	-3.0	1.9
member holonym	2.4	2.8	2.4	1.9	-2.4	2.9	-2.3	2.6	2.7	-2.4
member meronym	2.4	2.9	2.4	1.9	-2.3	-2.9	2.3	-2.5	-2.8	2.5
synset domain region of	-3.1	-0.3	3.1	-3.3	1.9	-0.9	2.0	-2.1	-1.2	1.0
member of domain region	-3.1	-0.3	3.2	-3.4	2.0	1.0	-2.1	2.2	1.3	-1.1
verb grou	3.5	3.4	3.3	-1.8	-2.8	0.0	-0.1	0.0	0.0	0.0
derivationally related forn	3.5	3.4	-3.2	3.4	3.2	0.0	0.0	-0.0	0.0	0.0



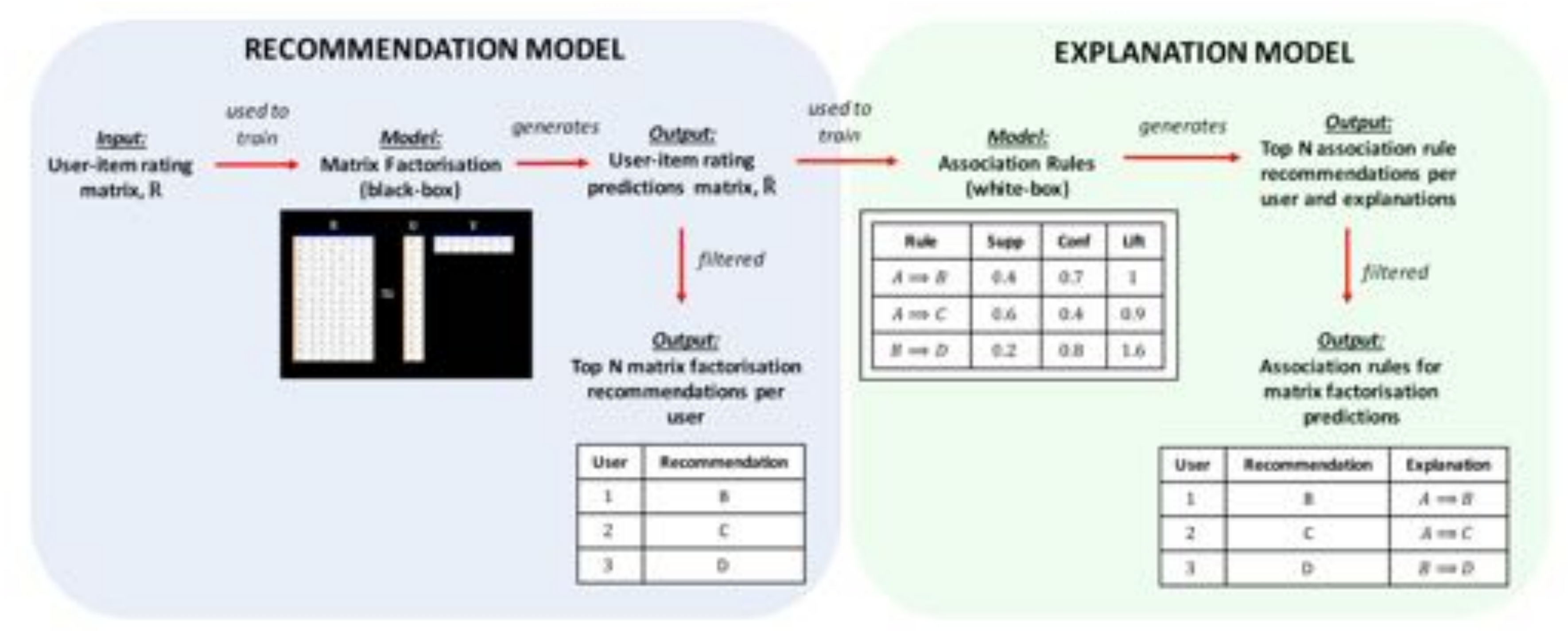
Predicates	Real Part					Imaginary Part				
musical arist	1.9	3.8	3.8	-1.7	-1.0	-2.5	0.4	-0.8	3.0	3.7
musical band	1.8	3.8	4.1	-1.8	-1.0	-2.5	0.3	-0.9	3.1	3.6
associated musical arist	3.7	3.2	3.7	3.4	3.3	0.7	0.1	0.2	-1.5	1.5
associated band	3.7	3.7	3.2	3.7	3.6	0.7	0.0	0.2	-1.5	1.5



Predicates	Real Part					Imaginary Part				
playsFor	3.6	-2.6	2.6	2.7	-3.1	2.5	3.0	2.8	2.6	-2.6
isAffiliatedTo	3.8	-2.6	2.6	2.6	-3.2	2.7	3.3	3.0	2.6	-2.8
hasNeighbor	0.9	2.5	2.9	3.5	2.2	0.0	-0.0	0.0	-0.1	-0.0
isMarriedTo	3.9	3.5	4.3	-2.1	0.0	0.0	-0.0	-0.0	0.0	0.0
isConnectedTo	-0.7	3.0	2.6	0.3	2.7	0.3	-0.1	-0.0	0.1	-0.0

Latent Feature Models - Post Hoc Interpretability

Generate an explanation model by training Bayesian Networks or Association Rules on the output of a Latent Feature Model. [Carmona et al. 2015, Peake et al. KDD 2018, Gusmão et al. 2018]



Rule	Supp	Conf	Lift
$A \Rightarrow B$	0.4	0.7	1
$A \Rightarrow C$	0.6	0.4	0.9
$B \Rightarrow D$	0.2	0.8	1.6

User	Recommendation
1	B
2	C
3	D

User	Recommendation	Explanation
1	B	$A \Rightarrow B$
2	C	$A \Rightarrow C$
3	D	$B \Rightarrow D$

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:

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

Neuro-Symbolic Reasoning

Neural and rule-based models have *complementary strengths and weaknesses*:

Neural Models

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

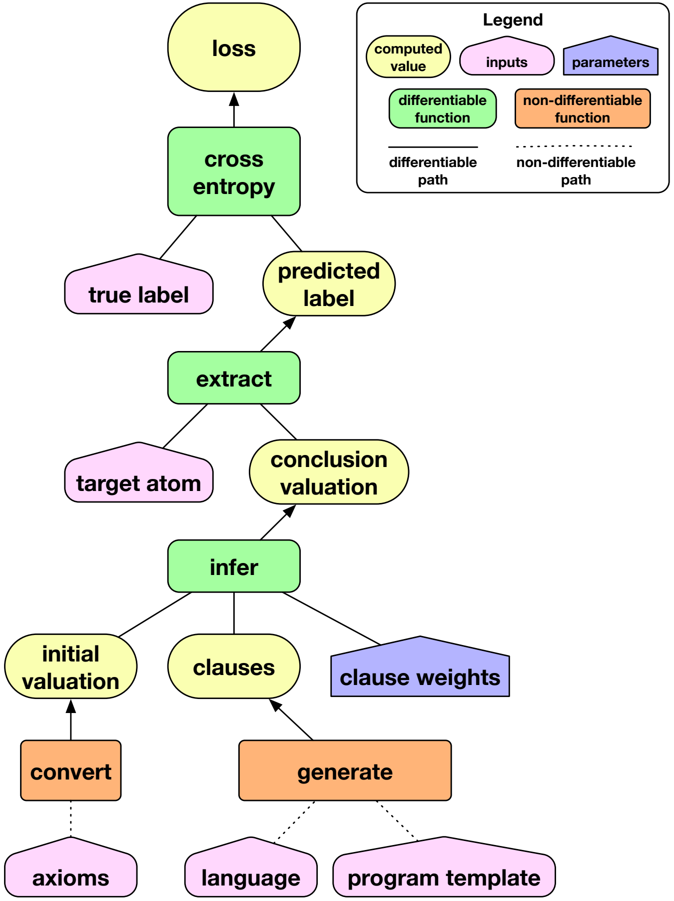
Rule-Based Models

- Can learn from small data
- Issues with high-dimensional, noisy, ambiguous inputs (*e.g.* images)
- Easy to **interpret**, provide **explanations**

Neuro-Symbolic Reasoning systems can combine the strengths of rule-based and neural architectures.

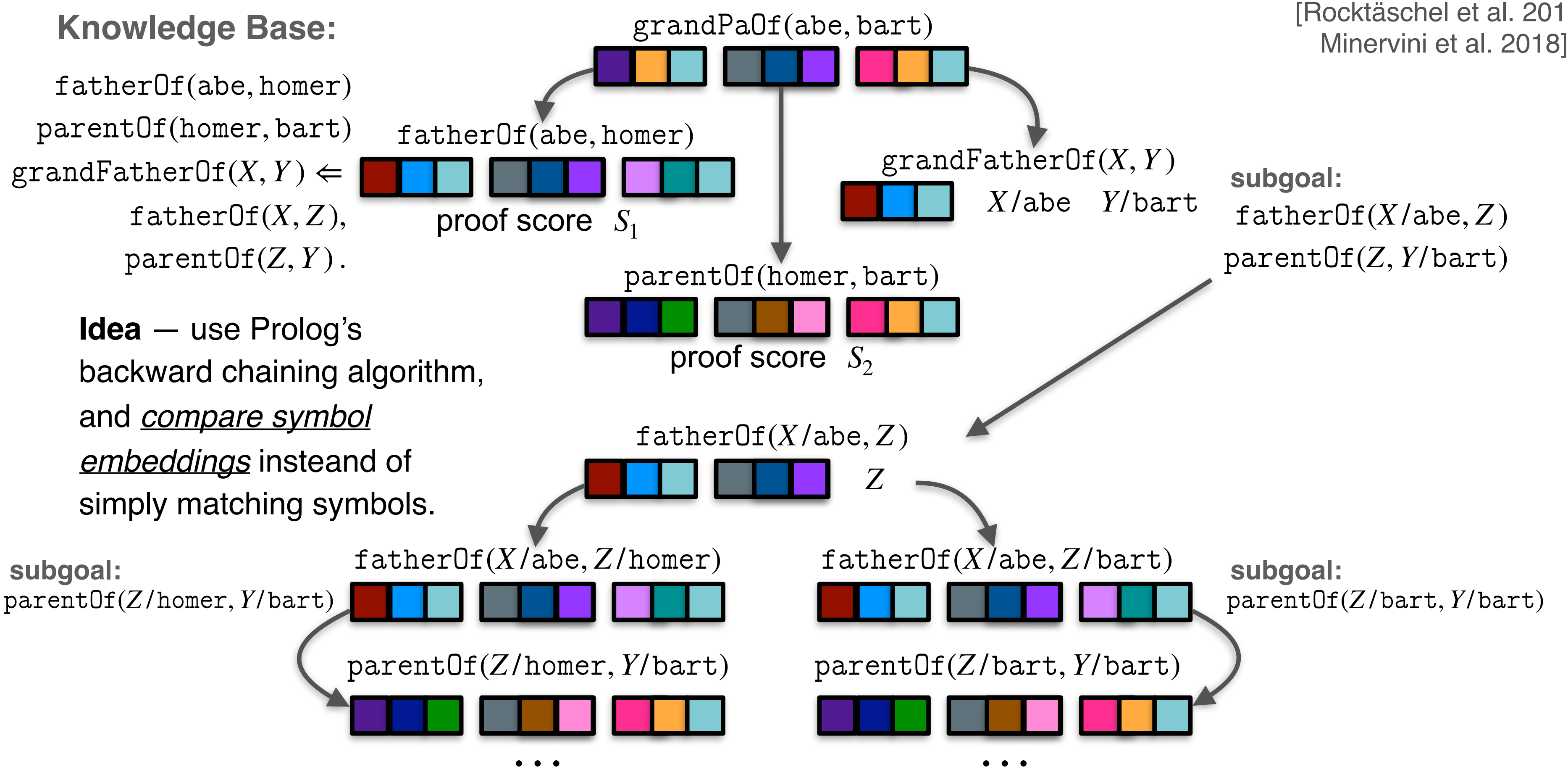
Forward Chaining — ∂ILP (Differentiable ILP) [Evans et al. JAIR 2018]

- Start with a **language definition** and a set of **background axioms**
- Generate a set of **clauses** — Datalog rules
- Given axioms and clauses, infer some **conclusions**
- Calculate the loss between the **reached** conclusions and the desired ones
- The system is **end-to-end differentiable**: we can back-propagate the error to the clause weights, representing our belief that rules should be in our program.



Backward Chaining — Differentiable Proving

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



Knowledge Base:

fatherOf(abe, homer)
parentOf(homer, bart)
grandFatherOf(X, Y) ⇐
fatherOf(X, Z),
parentOf(Z, Y).

Idea — use Prolog's backward chaining algorithm, and *compare symbol embeddings* instead of simply matching symbols.

grandPaOf(abe, bart)

fatherOf(abe, homer)

proof score S_1

parentOf(homer, bart)

proof score S_2

grandFatherOf(X, Y)

X/abe Y/bart

subgoal:
fatherOf(X/abe, Z)
parentOf(Z, Y/bart)

fatherOf(X/abe, Z)

Z

fatherOf(X/abe, Z/homer)

subgoal:
parentOf(Z/homer, Y/bart)

parentOf(Z/homer, Y/bart)

...

fatherOf(X/abe, Z/bart)

subgoal:
parentOf(Z/bart, Y/bart)

parentOf(Z/bart, Y/bart)

...

Differentiable Proving — Rule Learning

Knowledge Base:

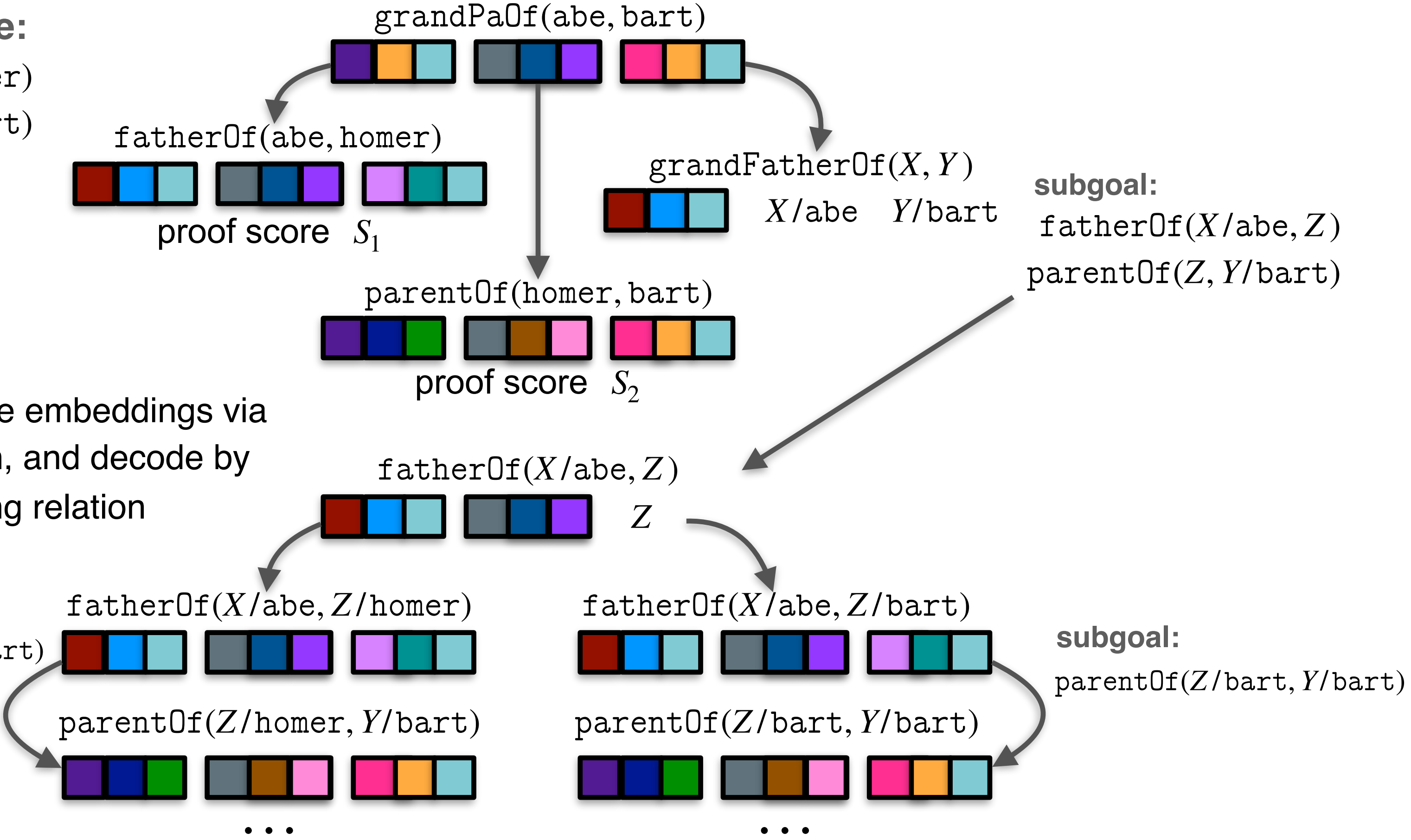
fatherOf(abe, homer)

parentOf(homer, bart)

$$\theta_1(X, Y) \Leftarrow \theta_2(X, Z), \theta_3(Z, Y).$$

$$\theta_1, \theta_2, \theta_3 \in \mathbb{R}^k$$

Idea — learn rule embeddings via backpropagation, and decode by looking at existing relation embeddings.



grandPaOf(abe, bart)

fatherOf(abe, homer)

proof score S_1

grandFatherOf(X, Y)

X/abe Y/bart

parentOf(homer, bart)

proof score S_2

subgoal:

fatherOf(X/abe, Z)

parentOf(Z, Y/bart)

fatherOf(X/abe, Z)

Z

subgoal:

parentOf(Z/homer, Y/bart)

fatherOf(X/abe, Z/homer)

parentOf(Z/homer, Y/bart)

. . .

fatherOf(X/abe, Z/bart)

subgoal:

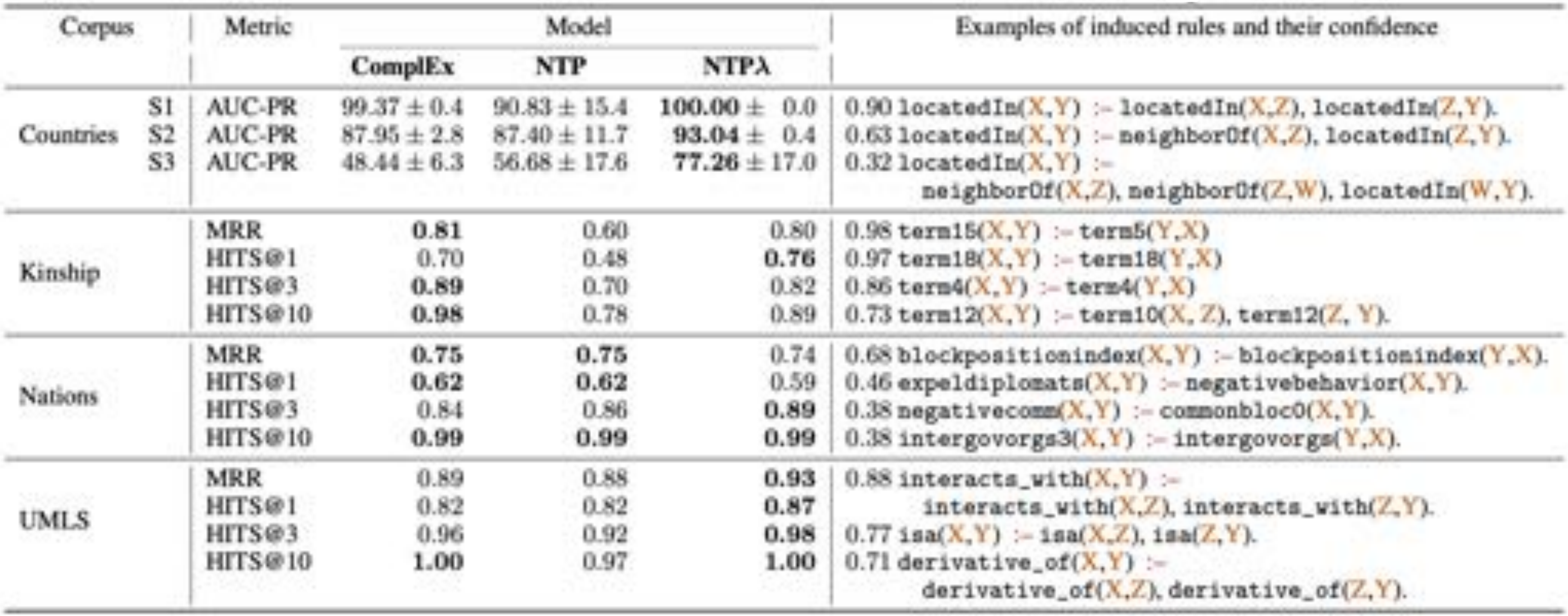
parentOf(Z/bart, Y/bart)

parentOf(Z/bart, Y/bart)

. . .

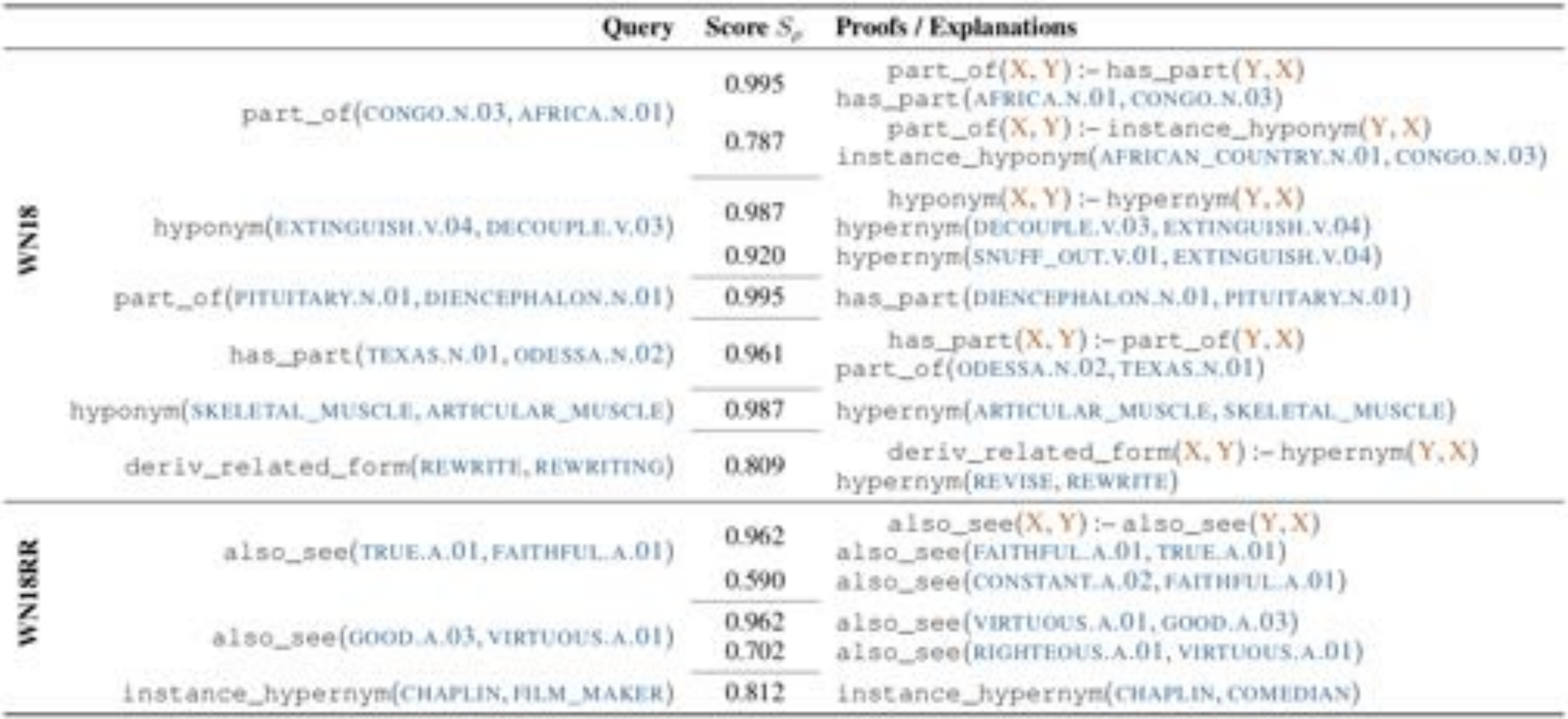
Differentiable Proving — Training

Train the model parameters — i.e. the entity and predicate embeddings, and the embeddings appearing in the rules — by *learning to prove* facts in the Knowledge Graph using all the remaining facts: $\mathscr{L}^{KB}(\theta) = -\sum_{FinK} \log \left[ntp_{\theta}^{KB \setminus F}(F, d) \right] - \sum_{\tilde{F} \sim corrupt(F)} \log \left[1 - ntp_{\theta}^{KB}(\tilde{F}, d) \right]$



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** ± 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** ± 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** ± 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**	
		HITS@3	0.96	0.92	**0.98**	0.77 isa(X,Y) :- isa(X,Z), isa(Z,Y).
		HITS@10	**1.00**	0.97	**1.00**	0.71 derivative_of(X,Y) :- derivative_of(X,Z), derivative_of(Z,Y).

Explainable Neural Link Prediction



	Query	Score S_p	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)
]

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.

Bibliography

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.

Bibliography

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)

Bibliography

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)

Bibliography

Pasquale Minervini, Matko Bosnjak, Tim Rocktäschel, Sebastian Riedel:
Towards Neural Theorem Proving at Scale. CoRR abs/1807.08204 (2018)

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

Applications

Luca Costabello
Accenture Labs
@lukostaz

Freddy Lécué
Inria, France
CortAIx@Thales, Canada
@freddylecue

Obstacle Identification Certification (Trust) - Transportation

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

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

XAI Technology: Deep learning and Epistemic uncertainty

Explainable On-Time Performance - Transportation

KLM / Transavia Flight Delay Prediction

PLANE INFO	ARRIVAL				TURNAROUND				DEPARTURE			
Status / Aircraft	Flight	ETA	Status	Delay Code	Gate	Slot	Progress	Milestones	Flight	ETA	Status	Delay Code
✓ urtwst	4567	18:30	Scheduled	-	345345	1	🟩		5678	19:00	Scheduled	-
! idsfew	4567	18:30	Delayed	ABC, DEF, GHI	345345	1	🟥		5678	19:00	Delayed	ABC, DEF, GHI
✓ psajdb	4567	18:30	Scheduled	ABC, DEF, GHI	345345	1	🟩		5678	19:00	Scheduled	ABC, DEF, GHI
⊘ kahdbs	4567	-	Cancelled	ABC, DEF, GHI	-	-			5678	-	Cancelled	ABC, DEF, GHI
! awwdfx	4567	18:35	Delayed	ABC, DEF, GHI	345345	1	🟨		5678	19:00	Delayed	ABC, DEF, GHI
! pdtsgs	4567	18:30	Delayed	ABC, DEF, GHI	345345	1	🟧		5678	19:00	Scheduled	ABC, DEF, GHI
✓ aedbac	4567	18:30	Scheduled	ABC, DEF, GHI	345345	1	🟩		5678	19:00	Scheduled	ABC, DEF, GHI
✓ aedbac	4567	18:30	Scheduled	ABC, DEF, GHI	345345	1	🟩		5678	19:00	Scheduled	ABC, DEF, GHI
✓ aedbac	4567	18:30	Scheduled	ABC, DEF, GHI	345345	1	🟩		5678	19:00	Scheduled	ABC, DEF, GHI
✓ aedbac	4567	18:30	Scheduled	ABC, DEF, GHI	345345	1	🟩		5678	19:00	Scheduled	ABC, DEF, GHI
✓ aedbac	4567	18:30	Scheduled	ABC, DEF, GHI	345345	1	🟩		5678	19:00	Scheduled	ABC, DEF, GHI
✓ aedbac	4567	18:30	Scheduled	ABC, DEF, GHI	345345	1	🟩		5678	19:00	Scheduled	ABC, DEF, GHI
✓ aedbac	4567	18:30	Scheduled	ABC, DEF, GHI	345345	1	🟩		5678	19:00	Scheduled	ABC, DEF, GHI
✓ aedbac	4567	18:30	Scheduled	ABC, DEF, GHI	345345	1	🟩		5678	19:00	Scheduled	ABC, DEF, GHI
✓ aedbac	4567	18:30	Scheduled	ABC, DEF, GHI	345345	1	🟩		5678	19:00	Scheduled	ABC, DEF, GHI

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

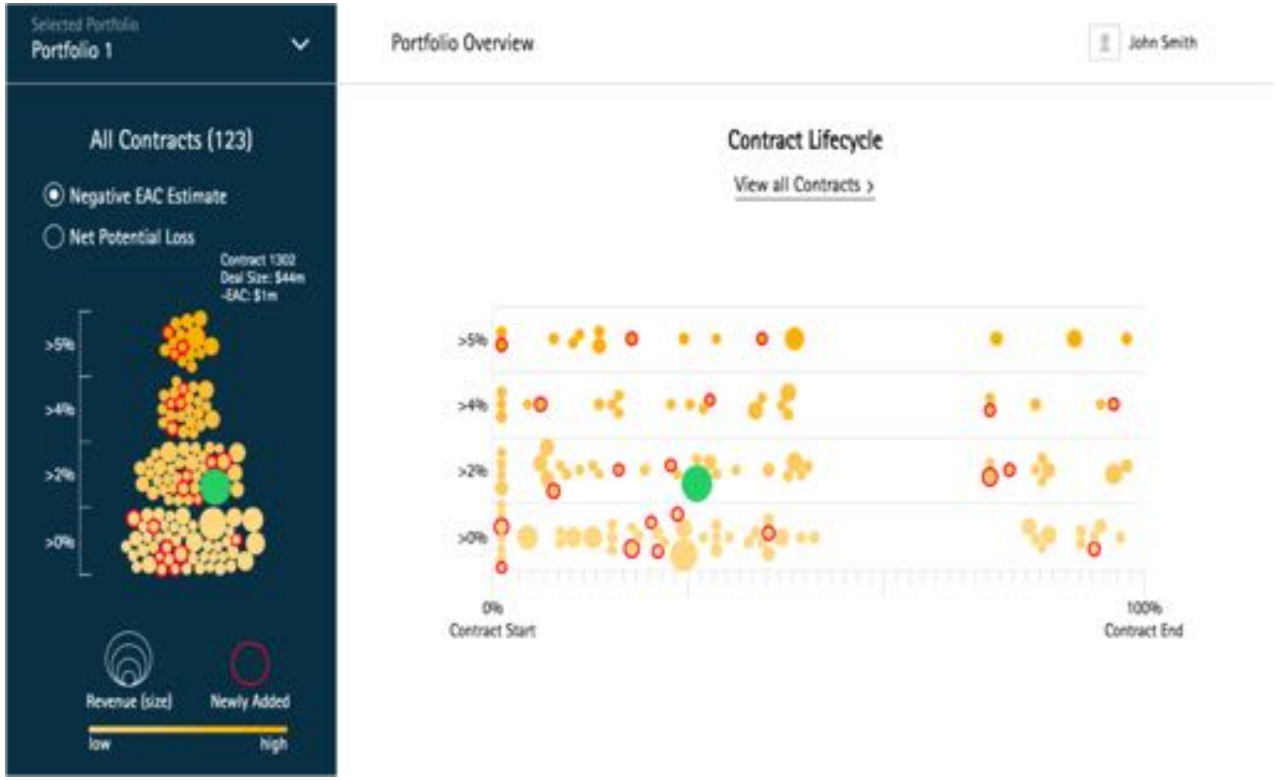
Nicholas McCarthy, Mohammad Karzand, Freddy Lecue: Amsterdam to Dublin Eventually Delayed? LSTM and Transfer Learning for Predicting Delays of Low Cost Airlines: AAAI 2019

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

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

XAI Technology: Knowledge graph embedded Sequence Learning using LSTMs

Explainable Risk Management - Finance



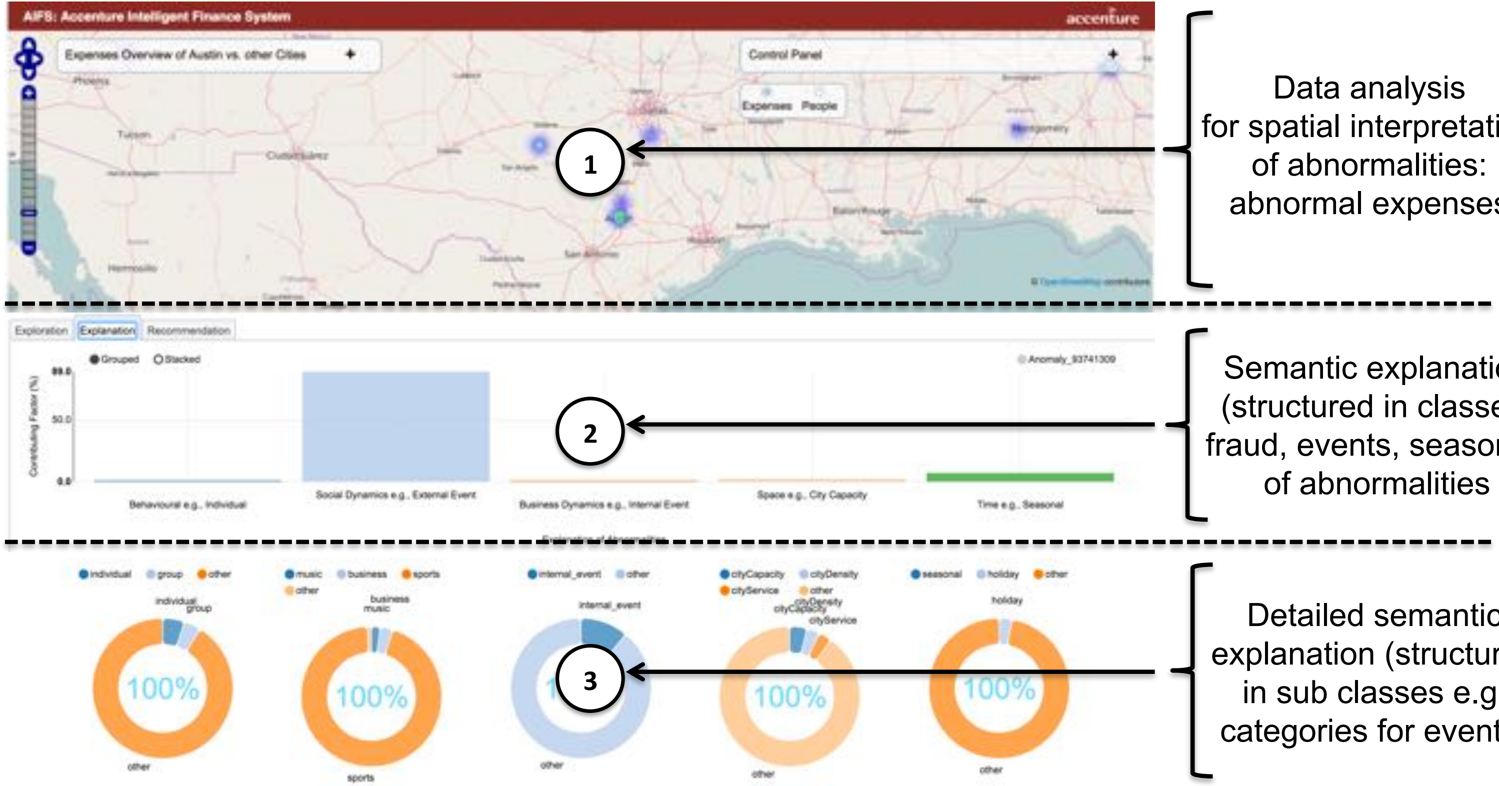
Jiewen Wu, Freddy Lécué, Christophe Guéret, Jer Hayes, Sara van de Moosdijk, Gemma Gallagher, Peter McCanney, Eugene Eichelberger: Personalizing Actions in Context for Risk Management Using Semantic Web Technologies. International Semantic Web Conference (2) 2017: 367-383

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

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

XAI Technology: Knowledge graph embedded Random Forrest

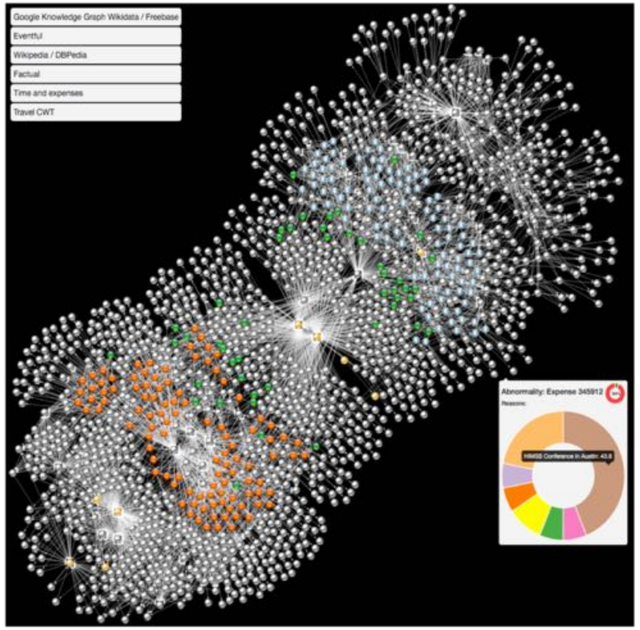
Explainable anomaly detection – Finance (Compliance)



Data analysis
for spatial interpretation
of abnormalities:
abnormal expenses

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

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



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

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

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

XAI Technology: Knowledge graph embedded Ensemble Learning

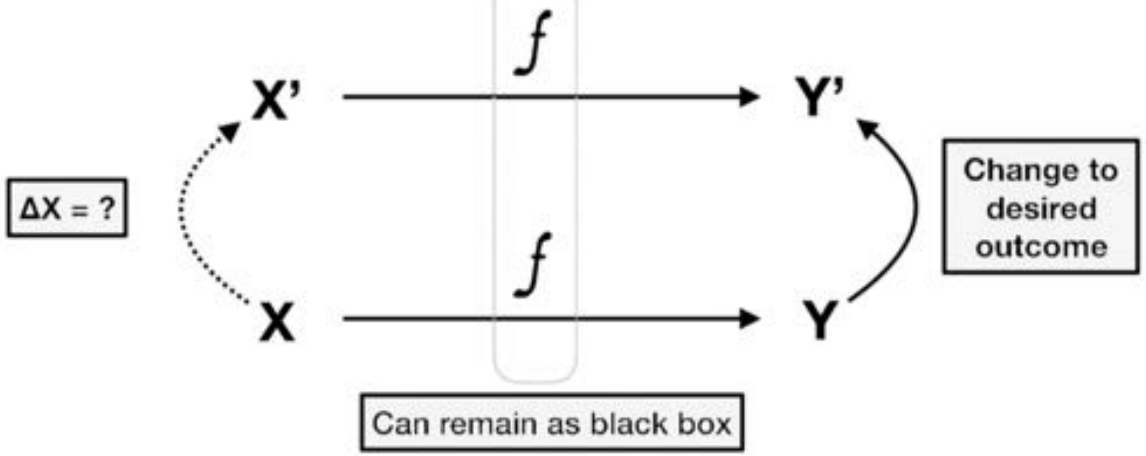
Counterfactual Explanations for Credit Decisions

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

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

AI Technology: Supervised learning, binary classification.

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



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

Counterfactual Explanations for Credit Decisions

Sorry, your loan application has been rejected.

Our analysis:

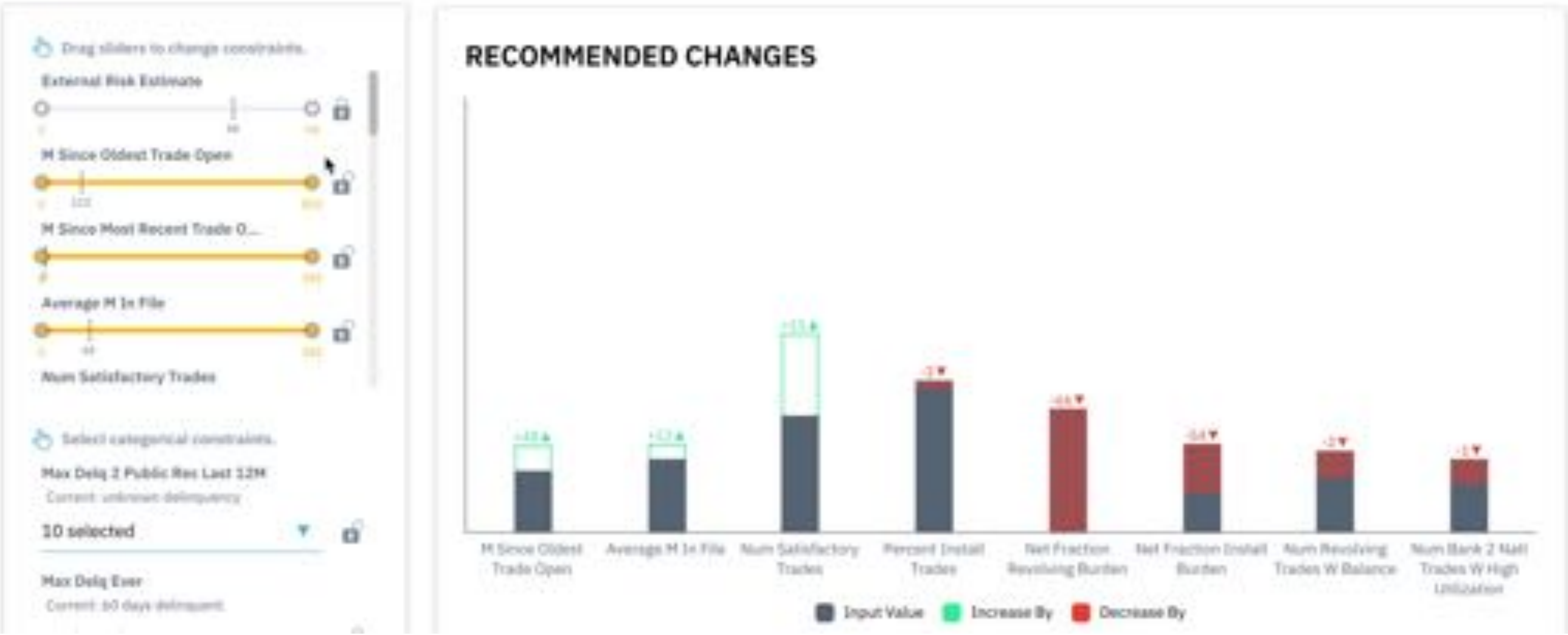
The following features were too high:

The following features were too low:

The following features require changes:

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

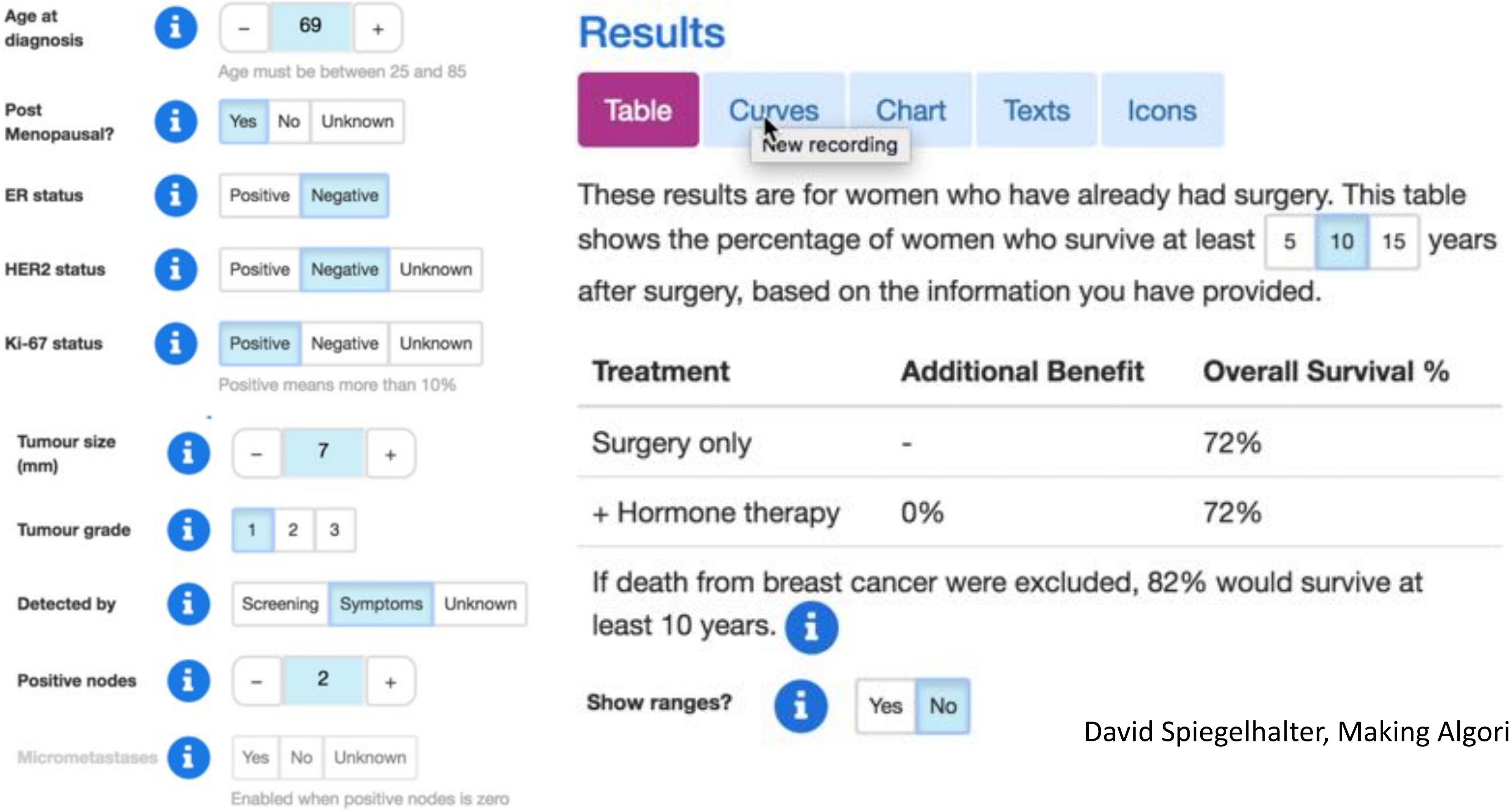
Rory Mc Grath, Luca Costabello, Chan Le Van, Paul Sweeney, Farbod Kamiab, Zhao Shen, Freddy Lécué: Interpretable Credit Application Predictions With Counterfactual Explanations. FEAP-AI4fin workshop, NeurIPS, 2018.



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

predict
breast cancer

Breast Cancer Survival Rate Prediction



Age at diagnosis – 69 +
Age must be between 25 and 85

Post Menopausal? [✓ Yes] No Unknown

ER status Positive [✓ Negative]

HER2 status Positive [✓ Negative] Unknown

Ki-67 status [✓ Positive] Negative Unknown
Positive means more than 10%

Tumour size (mm) – 7 +

Tumour grade [✓ 1] 2 3

Detected by Screening [✓ Symptoms] Unknown

Positive nodes – 2 +

Micrometastases Yes No Unknown
Enabled when positive nodes is zero

Results

These results are for women who have already had surgery. This table shows the percentage of women who survive at least 5 [✓ 10] 15 years after surgery, based on the information you have provided.

Treatment	Additional Benefit	Overall Survival %
Surgery only	-	72%
+ Hormone therapy	0%	72%

If death from breast cancer were excluded, 82% would survive at least 10 years.

Show ranges? Yes [✓ No]

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

AI Technology: competing risk analysis

XAI Technology: Interactive explanations, Multiple representations.

David Spiegelhalter, Making Algorithms trustworthy, NeurIPS 2018 Keynote

predict.nhs.uk/tool

(Some) Software Resources

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

Conclusions

Take-Home Messages

- Explainable AI is motivated by **real-world application of AI**

- Not a new problem – a reformulation of past research challenges in AI

- Multi-disciplinary: multiple AI fields, HCI, 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

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

Luca Costabello
Accenture Labs
@lukostaz

Fosca Giannotti
ISTI-CNR,
University of Pisa

Riccardo Guidotti
ISTI-CNR, University of Pisa
@rikdrive8s

Pascal Hitzler
Wright State University
@pascalhitzler

xaitutorial2019.github.io

Freddy Lécué
Inria, France
CortAIx@Thales, Canada
@freddylecue

Pasquale Minervini
University College London
@PMinervini

Kamruzzaman Sarker
Wright State University
@smkpallob