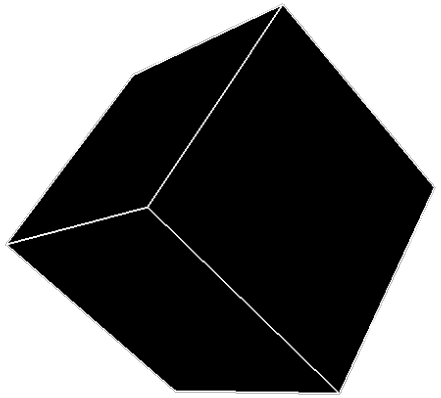
On Explainable AI:

From Theory to Motivation, Applications and Limitations

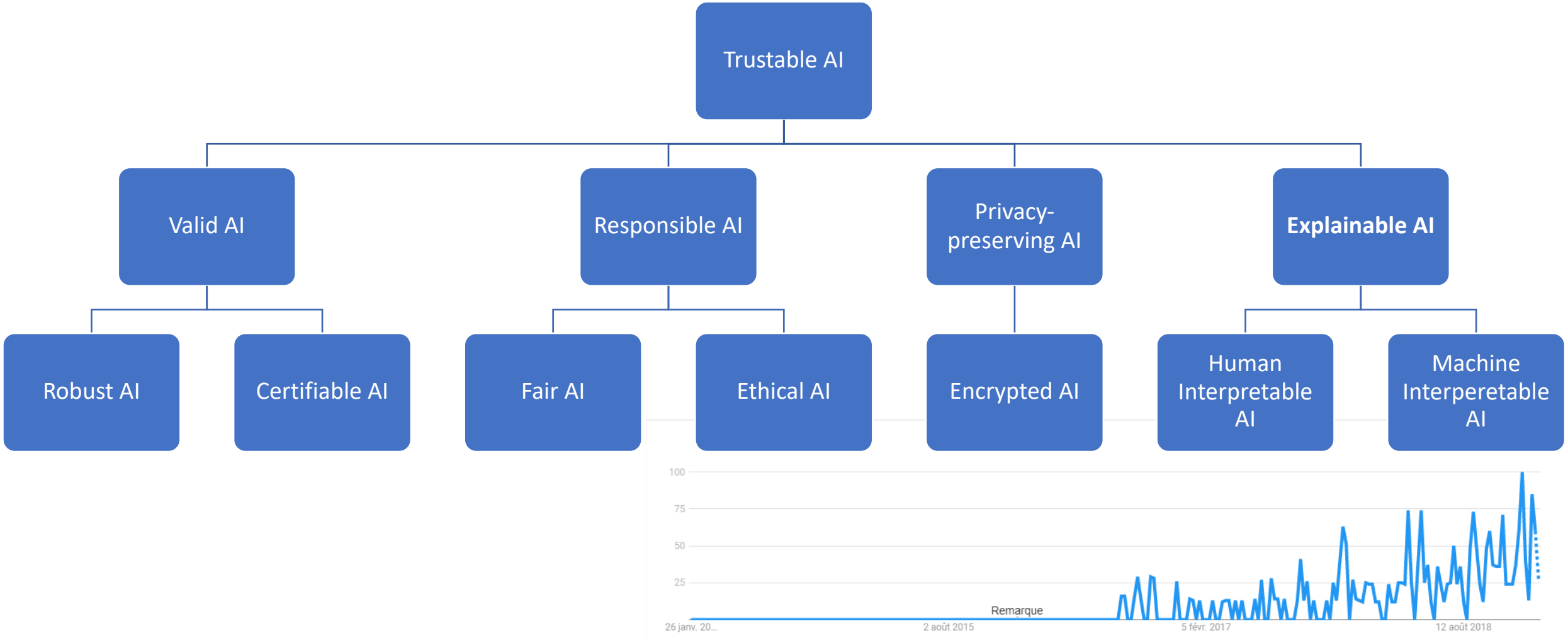
Freddy Lécué
Inria, France
CortAIx@Thales, Canada
@freddylecue

Pasquale Minervini
University College London
@PMinervini



Scope

*AI Context for Industrial Adoption



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 .

Outline

Tutorial Outline (1)

- **Explanation in Artificial Intelligence** **9:00 – 10:00**
  - Motivation
  - Definitions & Properties
  - Explanations in Different AI fields
  - The Role of Humans
  - Evaluation Protocols & Metrics
- **Explanation in Machine Learning** **10:00 – 11:00**
  - Explanation Taxonomy
  - Explanation in Machine Learning
- *Break* **11:00 – 11:30**

Tutorial Outline (2)

- **On the Role of Knowledge Graph in Explainable AI** **11:30 – 12:30**
  - Knowledge Graphs
  - Extending Machine Learning Systems with Knowledge Graphs
- *Break* **12:30 – 13:30**
- **On the Role of Reasoning in Explainable AI** **13:30 – 15:30**
  - Relational Learning
  - On Combining Neural Networks with Logic Programming
- *Break* **15:30 – 16:00**
- **Industrial Applications of XAI** **16:00 – 17:00**
- **Conclusion + Q&A** **17:00 – 18:00**

Motivation

Business to Customer

Gary Chavez added a photo you might be in. ...

about a minute ago ·

Critical Systems


BNSF

Markets We Serve (Critical Systems)

Aerospace

Space

Ground Transportation

Defence

Security

Trusted Partner For A Safer World

But not Only Critical Systems

COMPAS recidivism black bias

Opinion

OP-ED CONTRIBUTOR

When a Computer Program Keeps You in Jail

By Rebecca Wexler

June 13, 2017

Motivation (2)

Finance:

- Credit scoring, loan approval
- Insurance quotes

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

The Big Read Artificial intelligence + Add to myFT

Insurance: Robots learn the business of covering risk

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

Save

Oliver Ralph MAY 16, 2017 24

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

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.

Stanford MEDICINE | News Center

Email Tweet

Researchers say use of artificial intelligence in medicine raises ethical questions

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

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

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

Rich Caruana
Microsoft Research
rcaruana@microsoft.com

Yin Lou
LinkedIn Corporation
ylou@linkedin.com

Johannes Gehrke
Microsoft
johannes@microsoft.com

Paul Koch
Microsoft Research
paulkoch@microsoft.com

Marc Sturm
NewYork-Presbyterian Hospital
mas9161@nyp.org

Noémie Elhadad
Columbia University
noemie.elhadad@columbia.edu

Rich Caruana, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm, Noemie Elhadad: Intelligible Models for HealthCare: Predicting Pneumonia Risk and Hospital 30-day Readmission. KDD 2015: 1721-1730

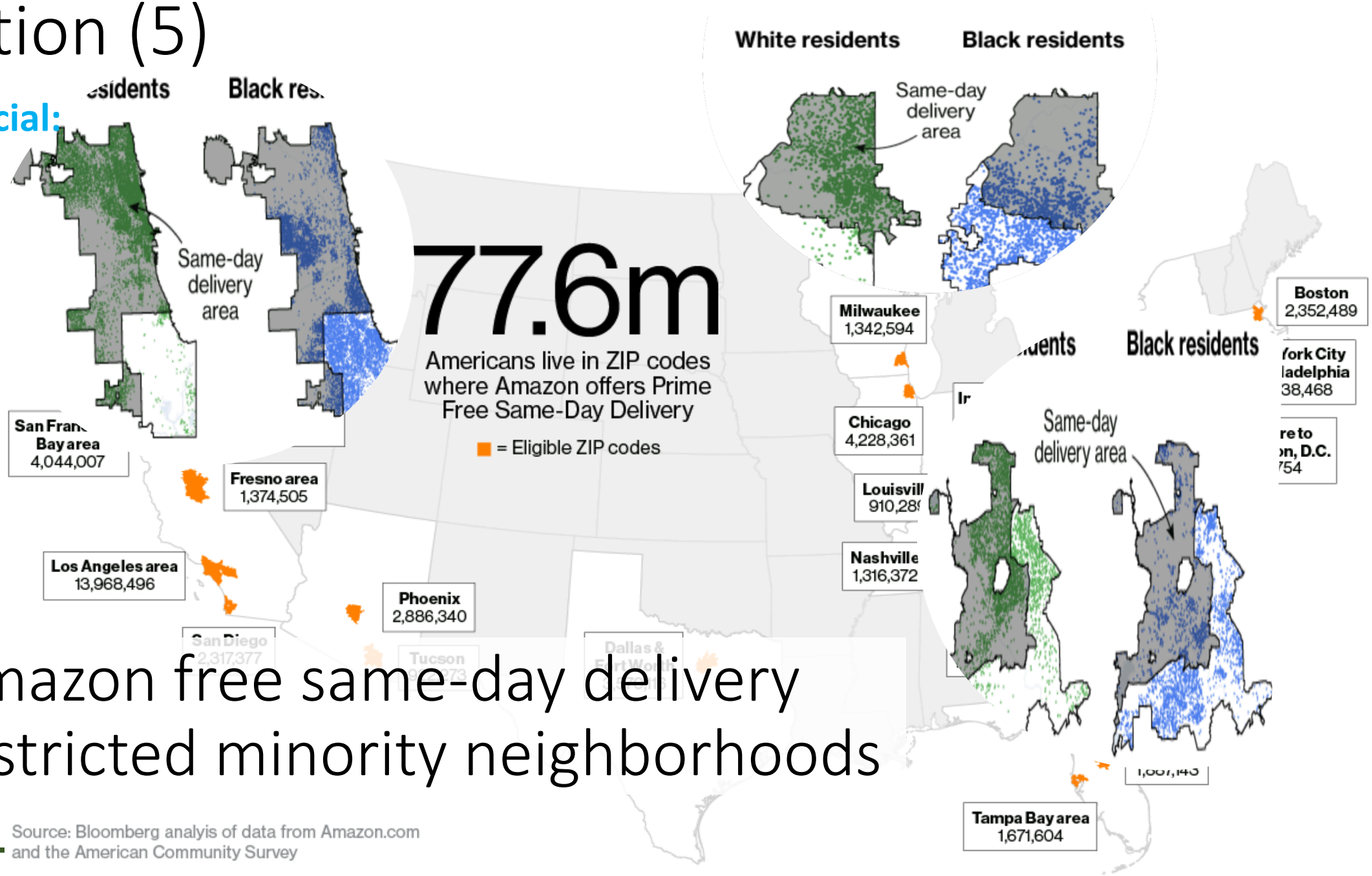
Motivation (4)

Human Resources – Talent Acquisition

- Discriminative Job Screening Software

Motivation (5)

Commercial:



No Amazon free same-day delivery for restricted minority neighborhoods

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

Industry Push for Explanation

Call for Explanation (1)

- User Acceptance & Trust

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

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

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

- Legal
  - Conformance to ethical standards, fairness
  - *Right to be informed*
  - Contestable decisions

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.

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

- Explanatory Debugging
  - Flawed performance metrics
  - Inadequate features
  - Distributional drift

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.

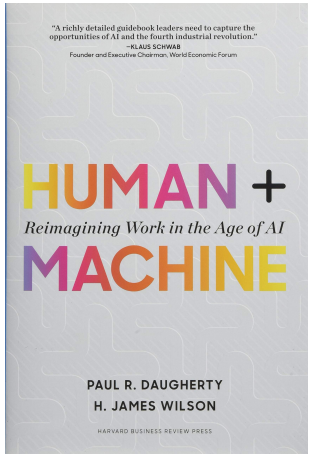
Weld, D., and Gagan Bansal. "The challenge of crafting intelligible intelligence." Communications of ACM (2018).

- Increase Insightfulness Informativeness
  - Uncovering causality

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

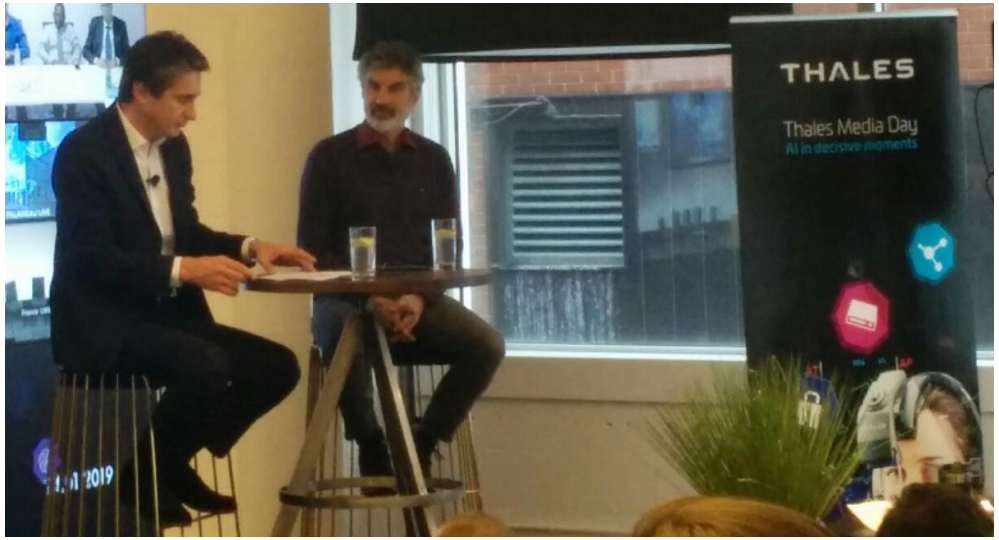
Judea Pearl: Causal Inference. NIPS Causality: Objectives and Assessment 2010: 39-58

Call for Explanation (2)



- **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!

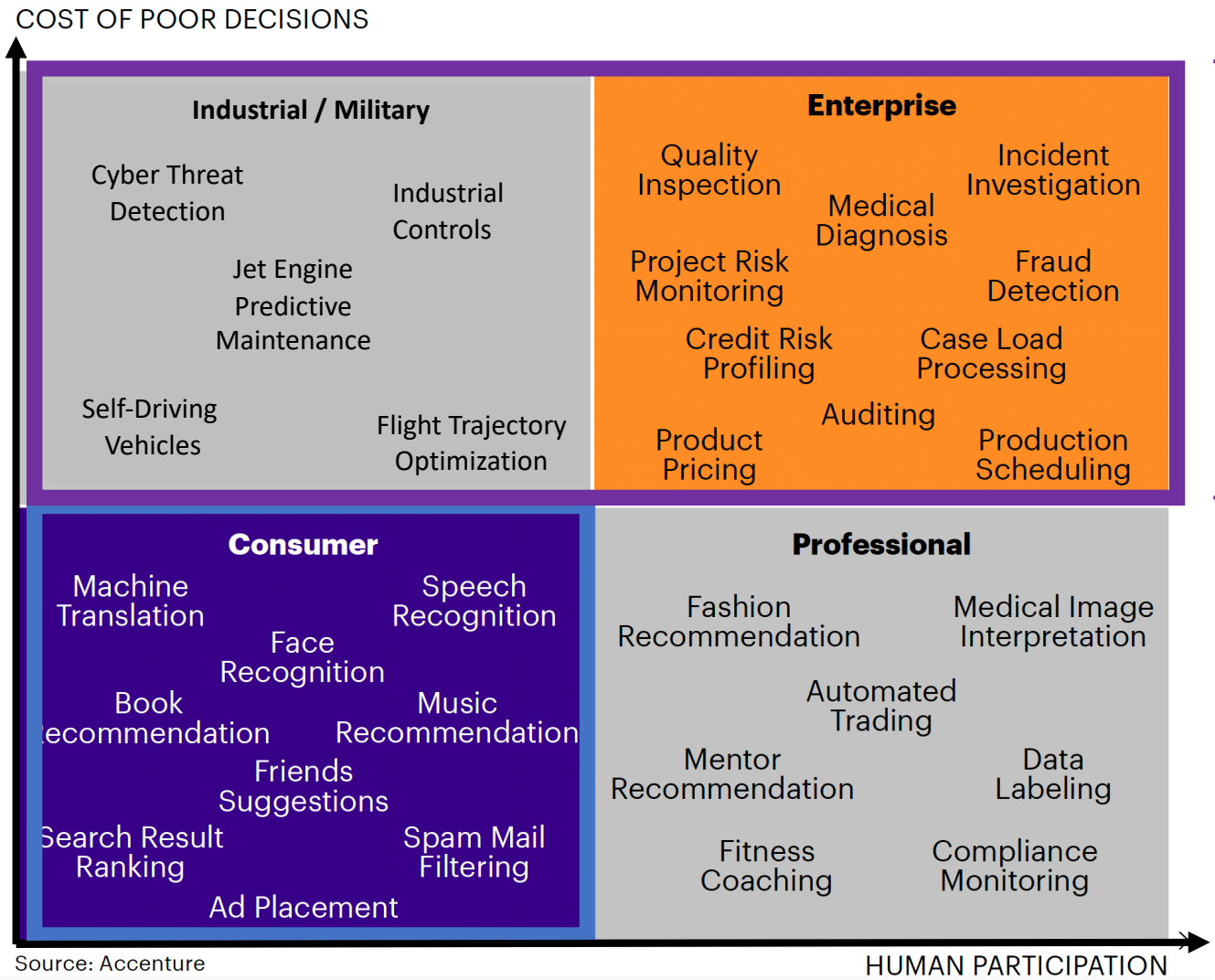
[Lepri et al. 2018]



Where is the Impact?

Trustable AI and eXplainable AI: a Reality Need

- The need for explainable AI rises with the potential cost of poor decisions



Most impactful successes of AI to come

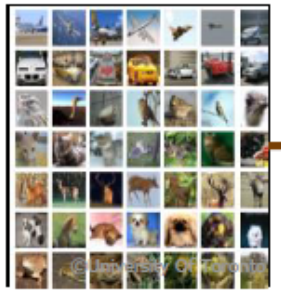
Most prominent successes of AI to date

Source: Accenture Point of View. Understanding Machines: Explainable AI. Freddy Lecue, Dadong Wan

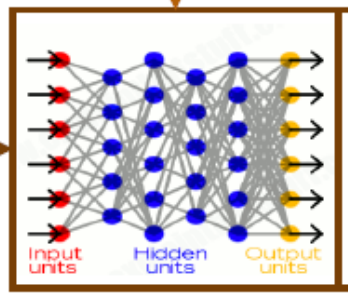
XAI in a Nutshell

XAI in a Nutshell

Today



Learning Process

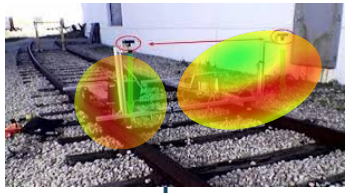


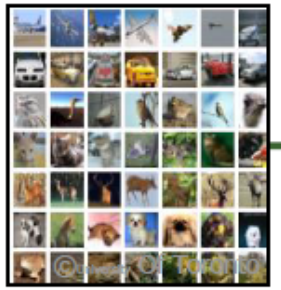
This is an obstacle on rail train

- Why did you do that?
- Why not something else?
- When do you succeed?
- When do you fail?
- When can I trust you?
- How do I correct an error?

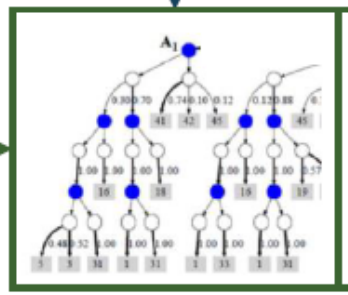
Training Data | Learned Function | Output | User with a Task

Tomorrow





New Learning Process



Obstacle on rail train

- **Obstruction covering full width**

- I understand why
- I understand why not
- I know when you'll succeed
- I know when you'll fail
- I know when to trust you
- I know why you erred

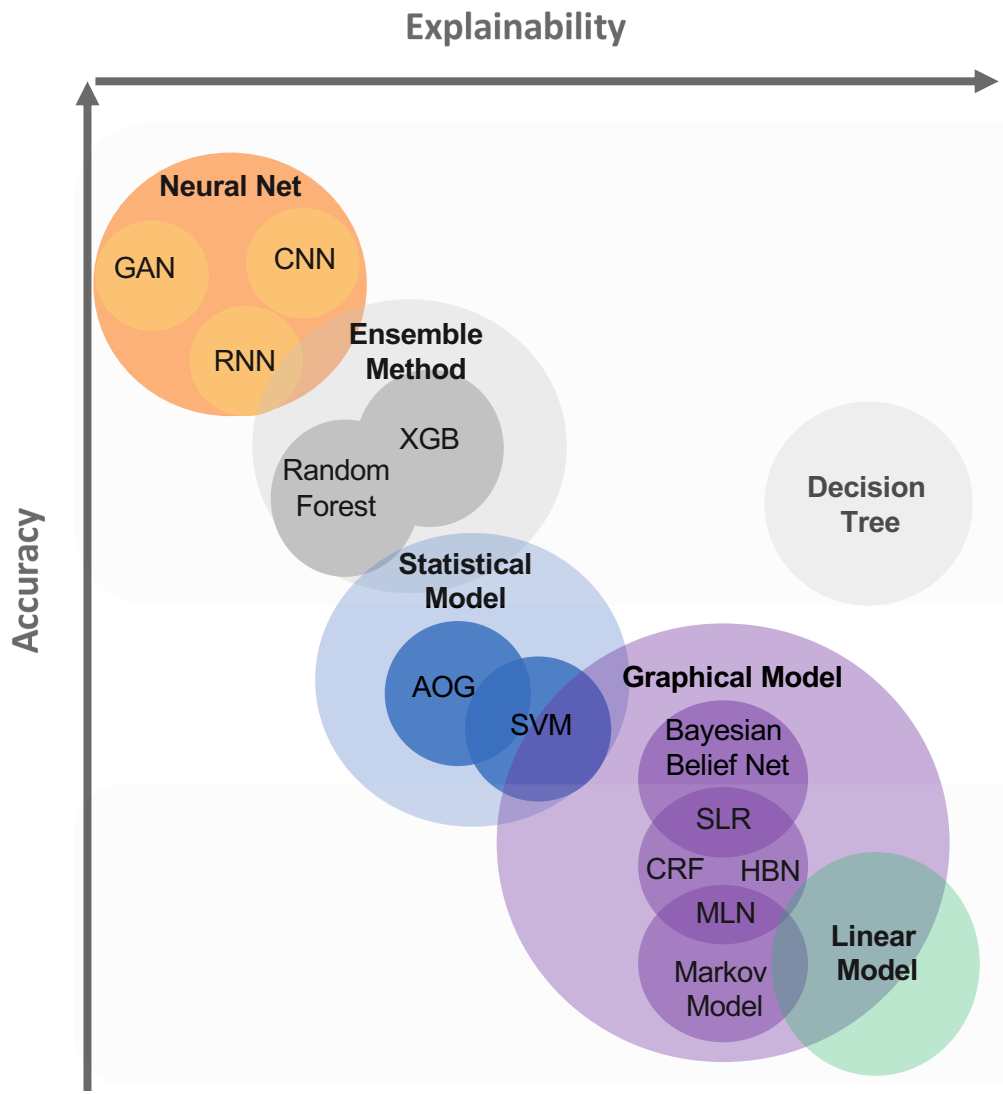
Training Data | Explainable Model | Explanation Interface | User with a Task

Source: https://www.cc.gatech.edu/~alanwags/DLAI2016/(Gunning)%20IJCAI-16%20DLAI%20WS.pdf

How to Explain? Accuracy vs. Explanability

Learning

- Challenges:
  - Supervised
  - Unsupervised learning
- Approach:
  - Representation Learning
  - Stochastic selection
- Output:
  - **Correlation**
  - **No causation**



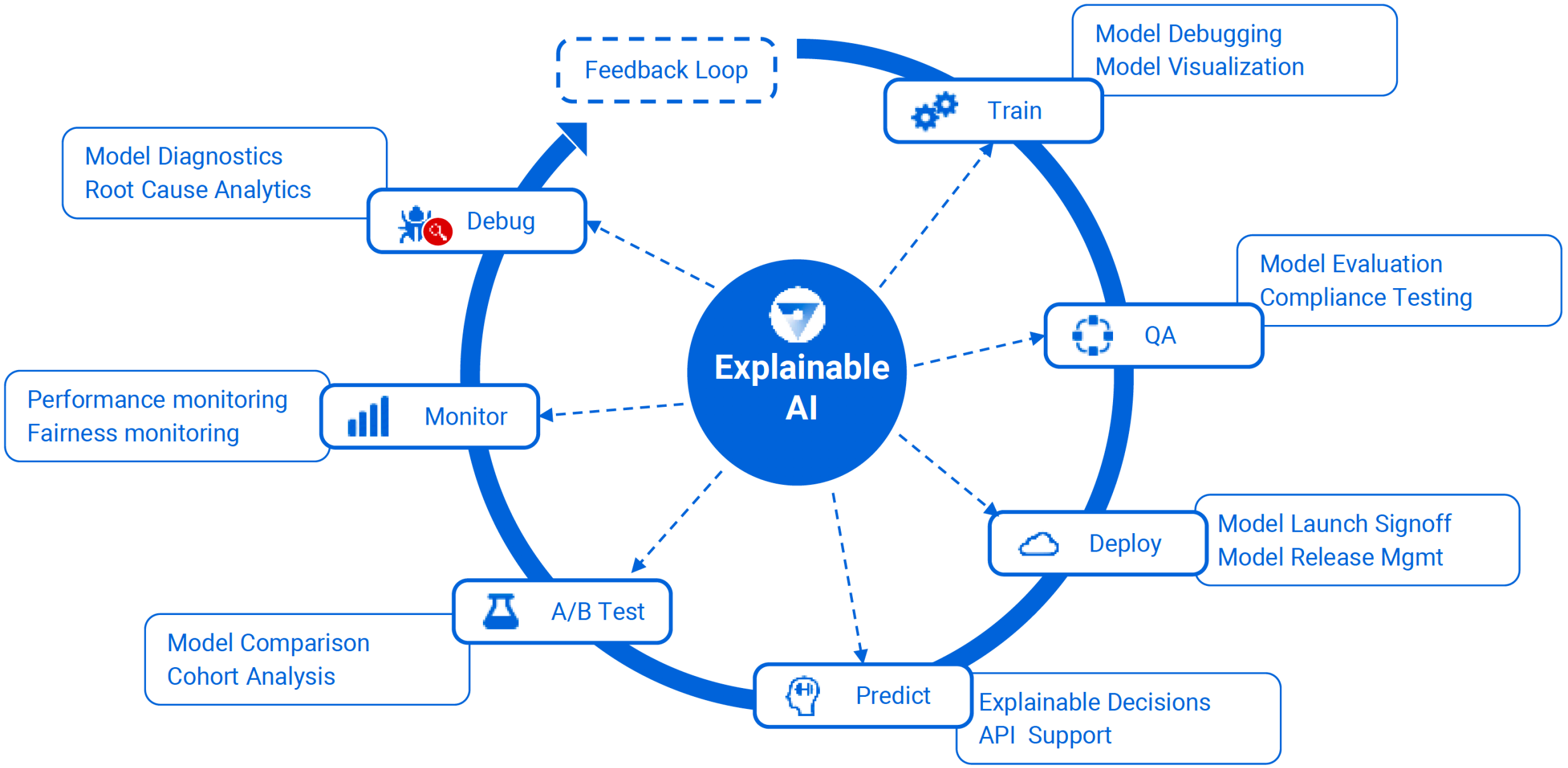
Interpretability

- Non-Linear functions
- Polynomial functions
- Quasi-Linear functions

XAI Objective

Supporting Industrialization of AI at Scale

Explainability by Design for AI Products



XAI Definitions

Explanation vs. Interpretability

Oxford Dictionary of English

explanation | ɛksplə'neɪʃ(ə)n |

noun

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

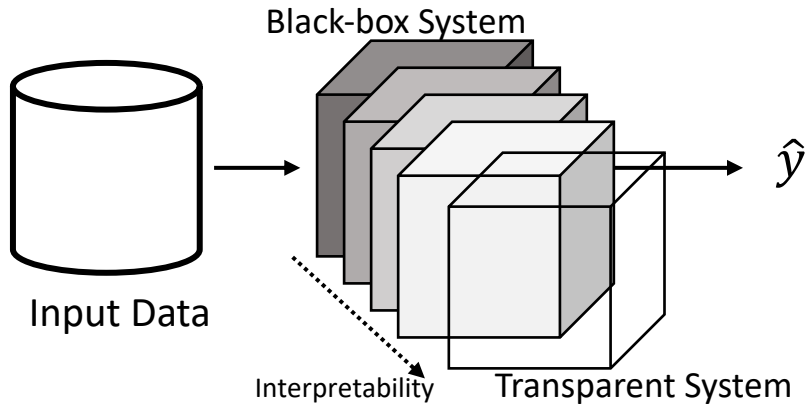
interpret | ɪn'tə:prɪt |

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

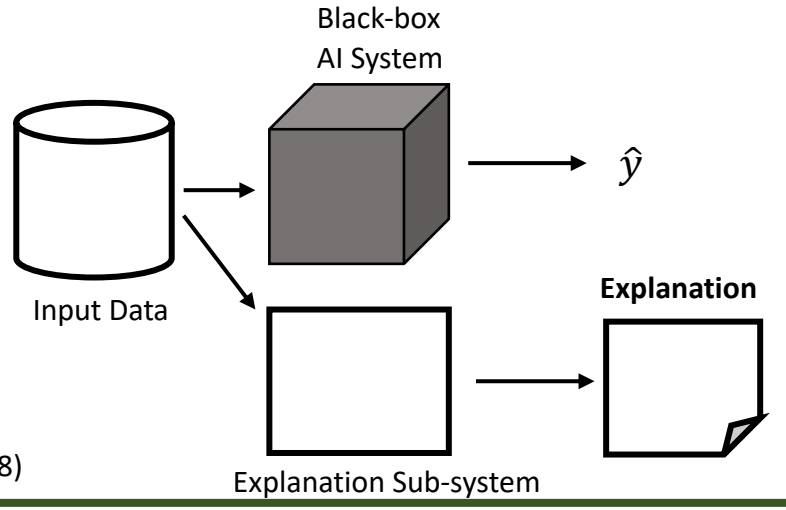
1 explain the meaning of (information or actions): *the evidence is difficult to interpret*.

Transparent Design vs Post-hoc Explanation

Transparent design reveals ***how*** **a model functions**.



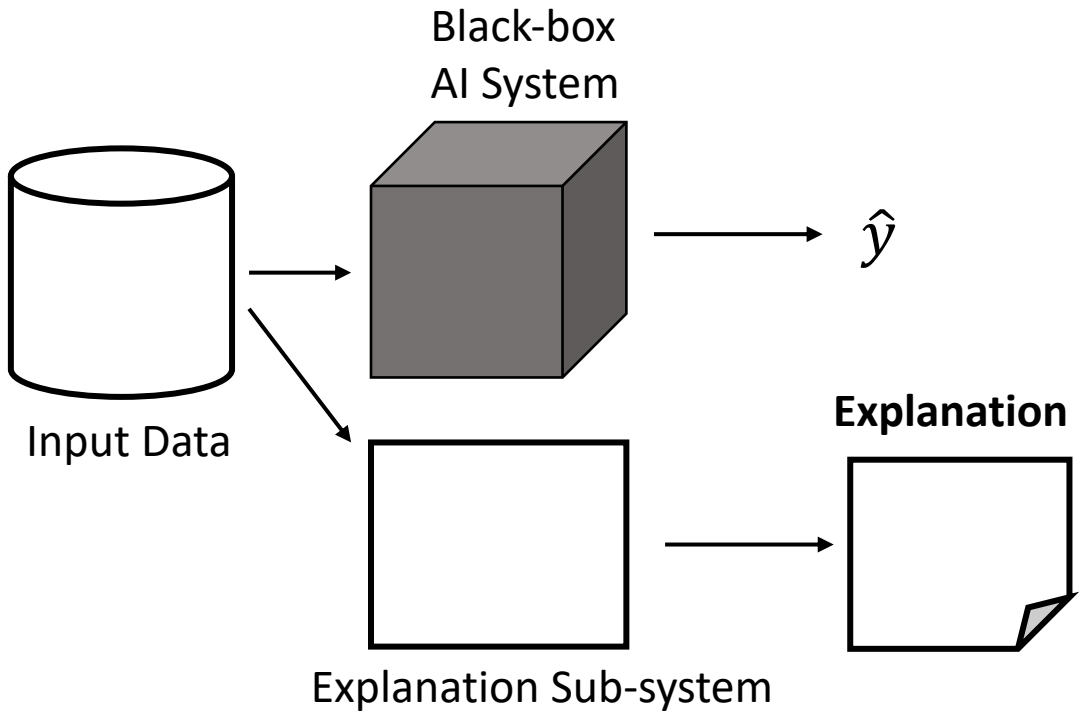
Post-hoc Explanation explains ***why*** **a black-box model behaved that way**.



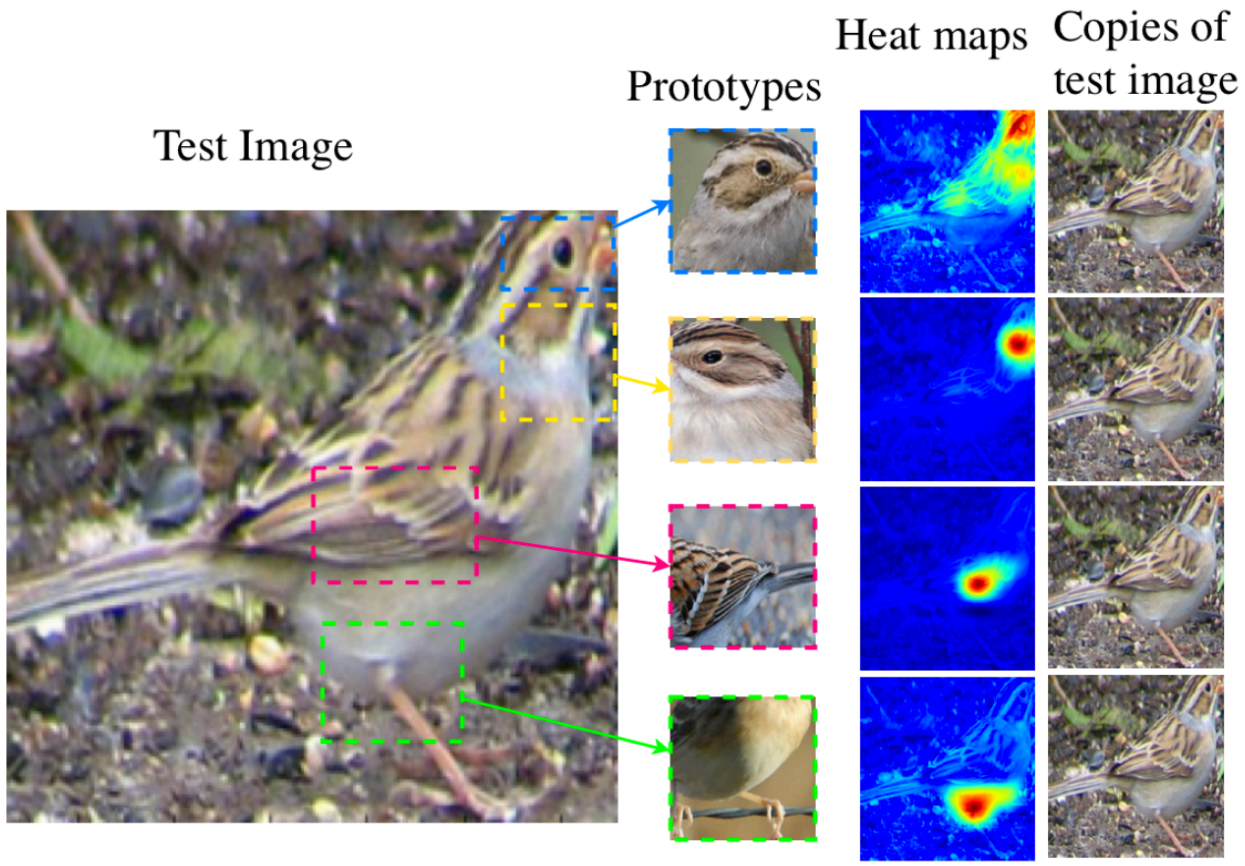
Brent D. Mittelstadt, Chris Russell, Sandra Wachter: Explaining Explanations in AI. CoRR abs/1811.01439 (2018)

So, What is an Explanation?

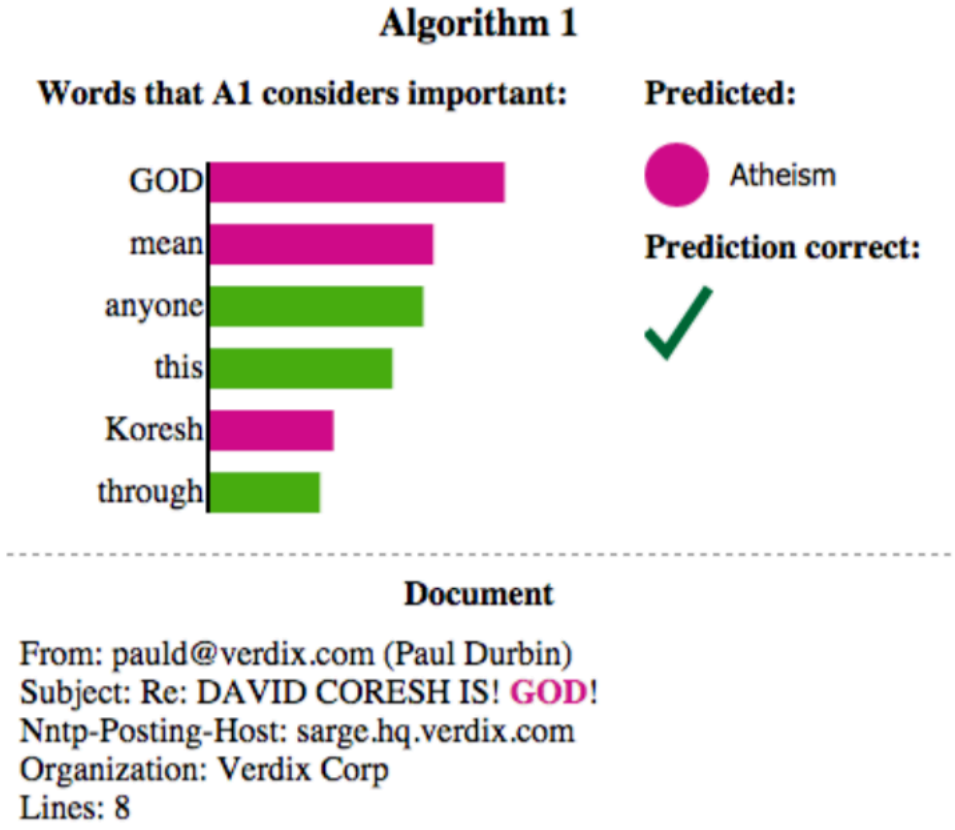
- **No formal, technical, agreed upon definition!**
- Comprehensive philosophical overview out of scope of the tutorial [Miller 2017]
- Not limited to machine learning! [Lipton 2016, Tomsett et al. 2018, Rudin 2018]



Examples



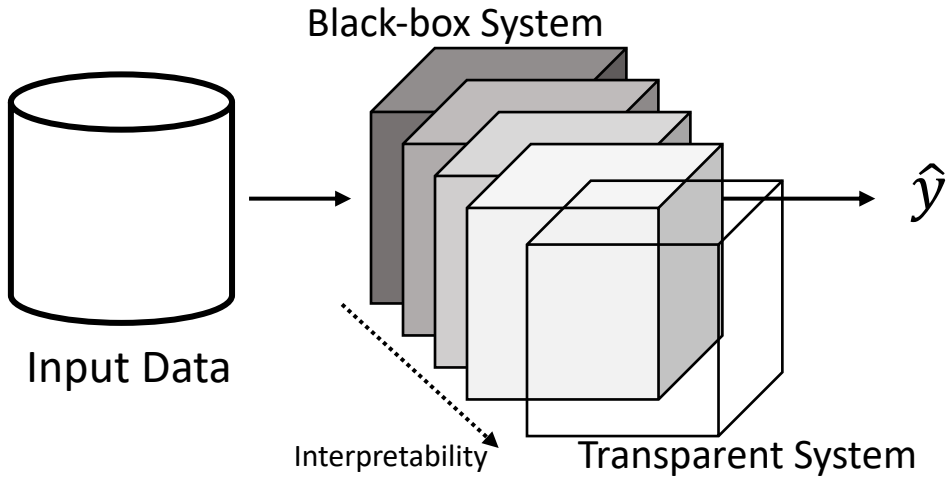
[Chen and Rudin 2018]



[Ribeiro et al. 2016]

What About Interpretability?

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



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

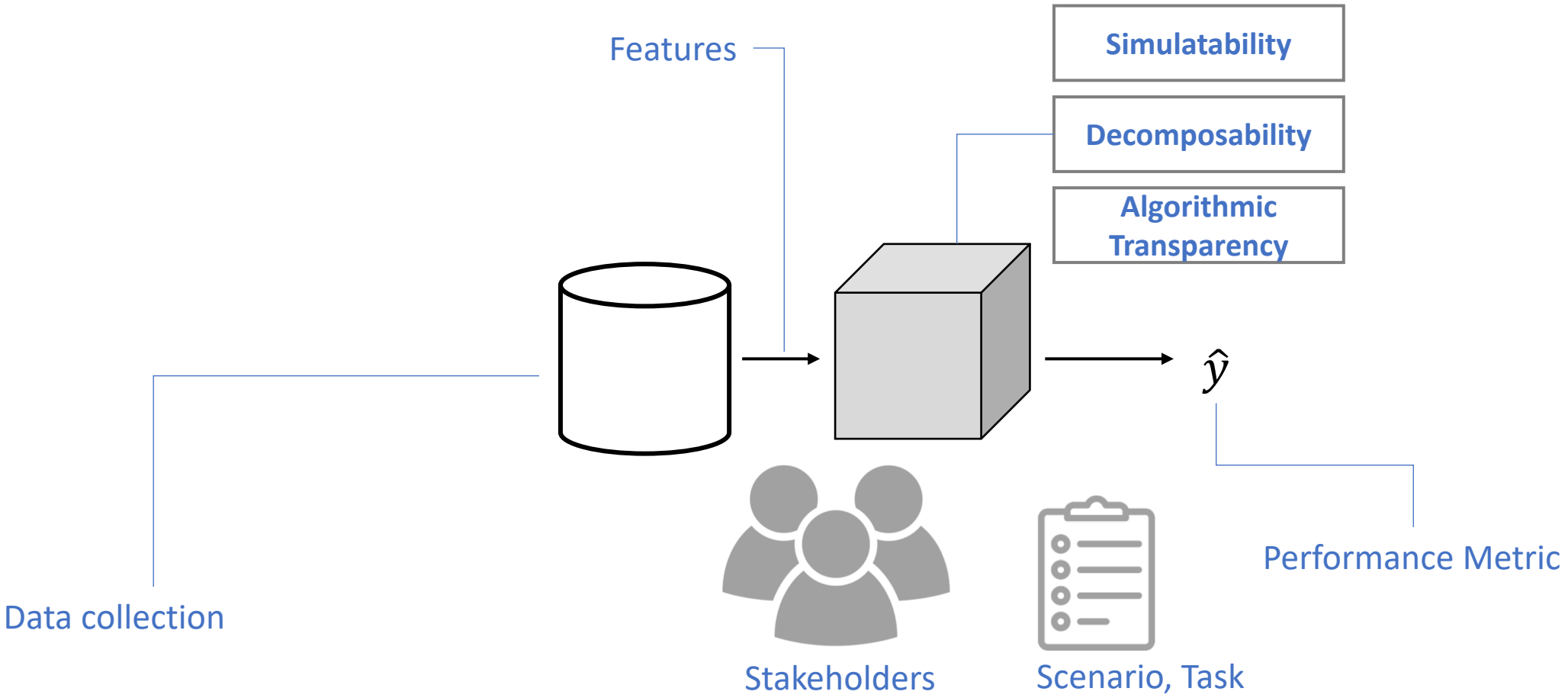
Algorithmic Transparency

Understanding at the level of **training algorithm**

Transparent Training Algorithm

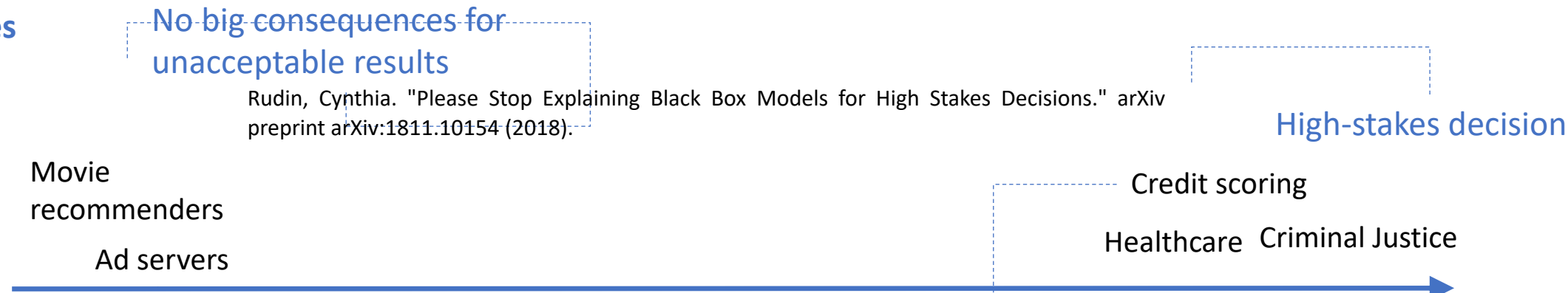
[Lipton 2016, Lepri et al. 2017, Mittelstadt et al. 2018, Weld and Bansal 2018]

Interpretability Goes Beyond the Model



Desire for Explainable AI **Must be Justified**

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



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

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

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

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

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

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.

On Role of Data In XAI

Interpretable Data for Interpretable Models

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

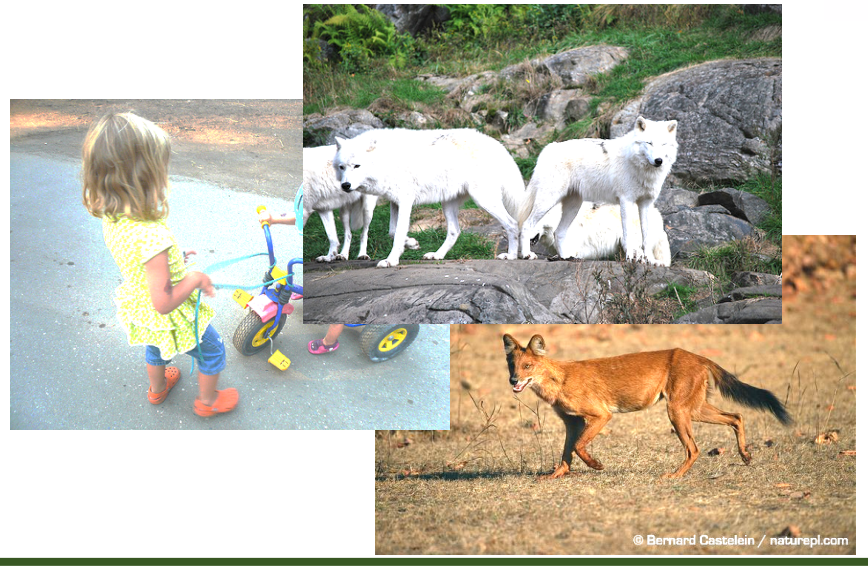
Field names

One row (4 fields)

2000 rows all told

Tabular

Images



Text

XAI Properties

(Some) Desired Properties of Explainable AI Systems (1)

- **Informativeness:** to which extend the model / prediction can be of use
- ***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.

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

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

(Some) Desired Properties of Explainable AI Systems (2)

- ***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.
- **Low cognitive load**: explanation should easy to understand

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.

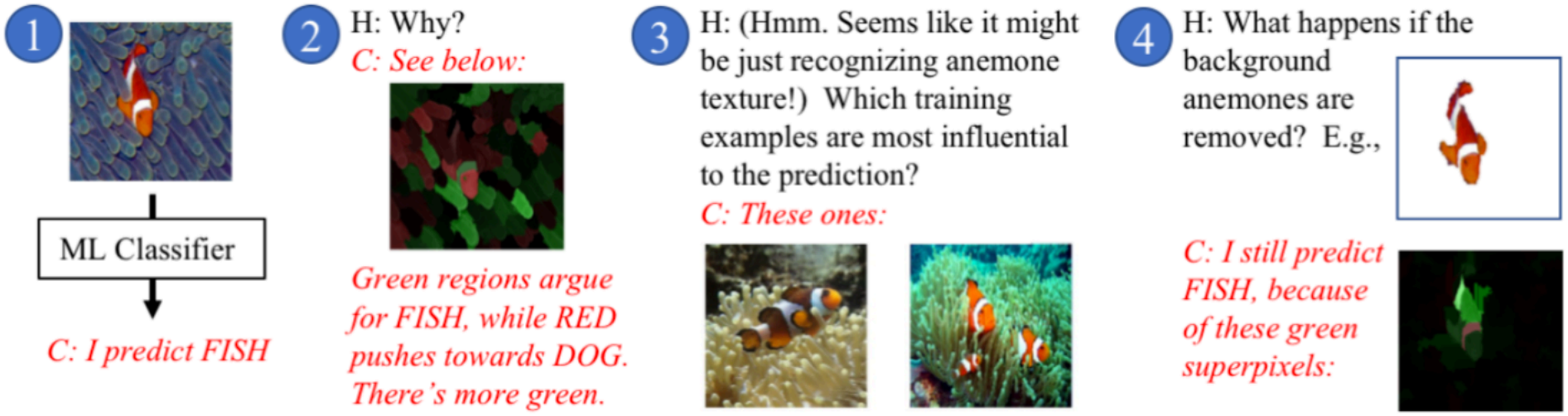
(Some) Desired Properties of Explainable AI Systems (3)

- ***Reliability and Robustness***: the interpretable model should maintain high levels of performance independently from small variations of the parameters or of the input data.
- ***Non-misleading:*** the interpretation sticks to the models, and do not hallucinate on behavior
- ***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.
- **Interactivity /Conversational:** explanation should be refined based on user profile, preference and experience

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

Mittelstadt, Brent, Chris Russell, and Sandra Wachter. "Explaining explanations in AI." arXiv preprint arXiv:1811.01439 (2018).

Explanation as *System-Human Conversation*



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

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

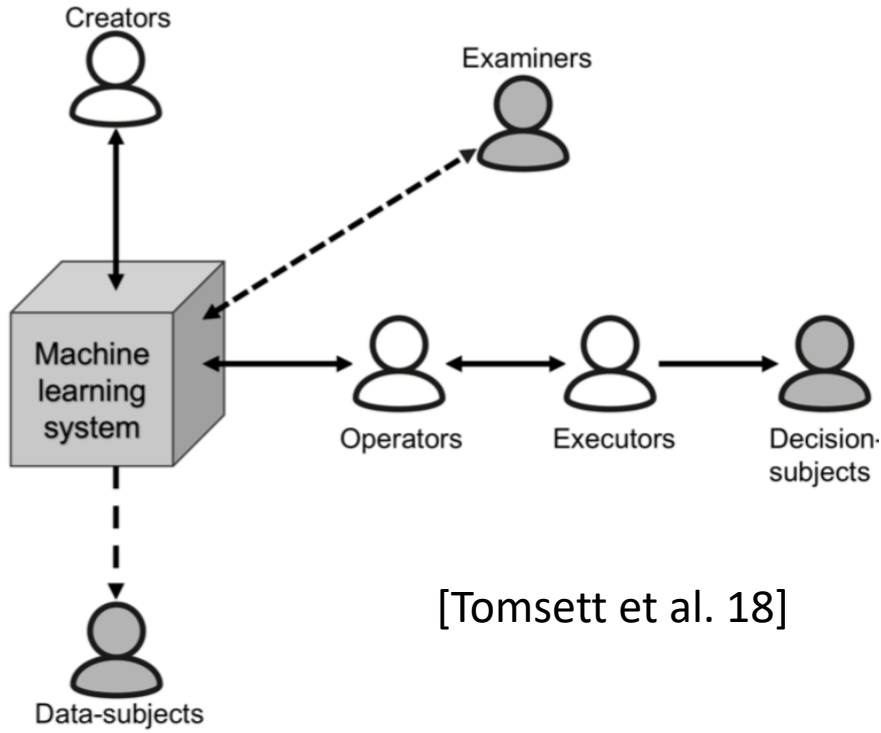
What about the Users?

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

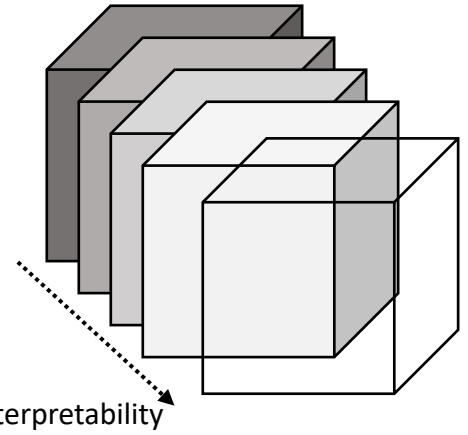
Doshi-Velez, Finale, and Been Kim. "Towards a rigorous science of interpretable machine learning." arXiv preprint arXiv:1702.08608 (2017).

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

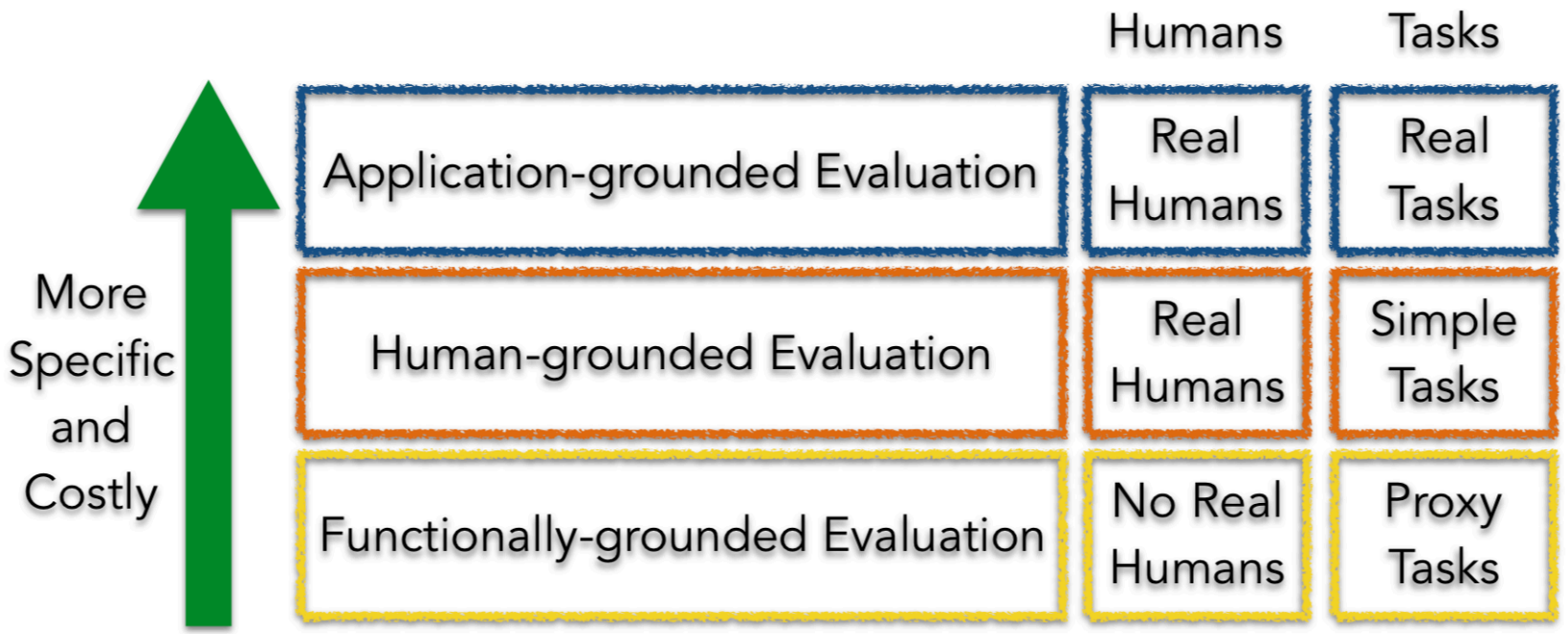
What about the Evaluation?

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



[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]

Human-based Evaluation for Feature Attribution-based Approaches

Have humans review attributions and/or compare them to (human provided) groundtruth on “feature importance”

Pros:

- Helps assess if attributions are human-intelligible
- Helps increase trust in the attribution method

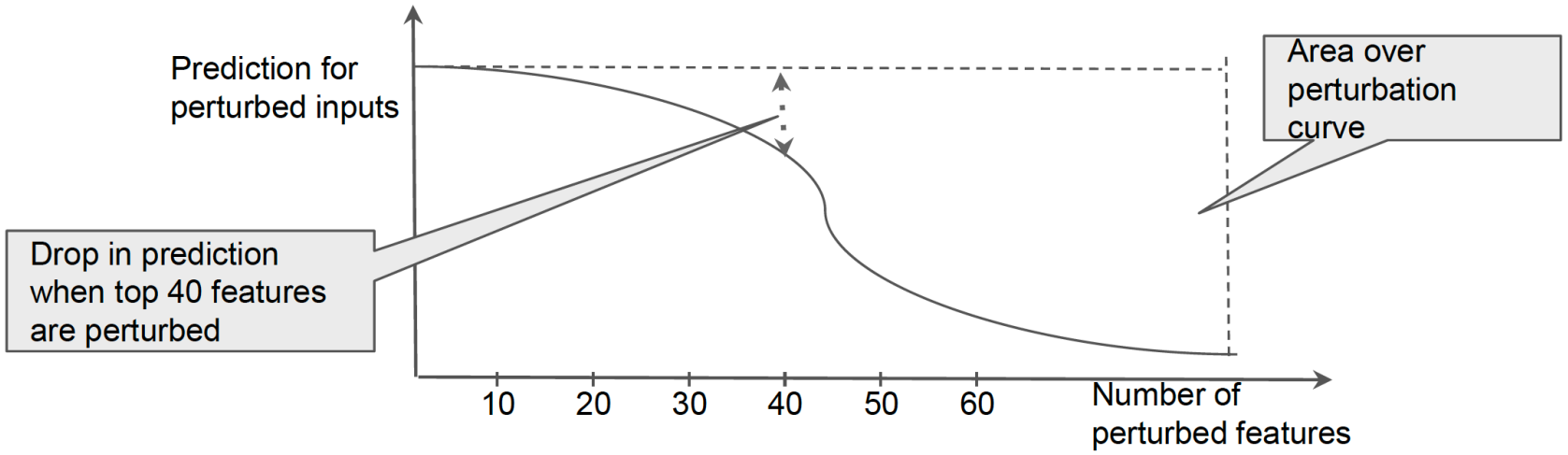
Cons:

- Attributions may appear incorrect because model reasons differently
- **Confirmation bias**

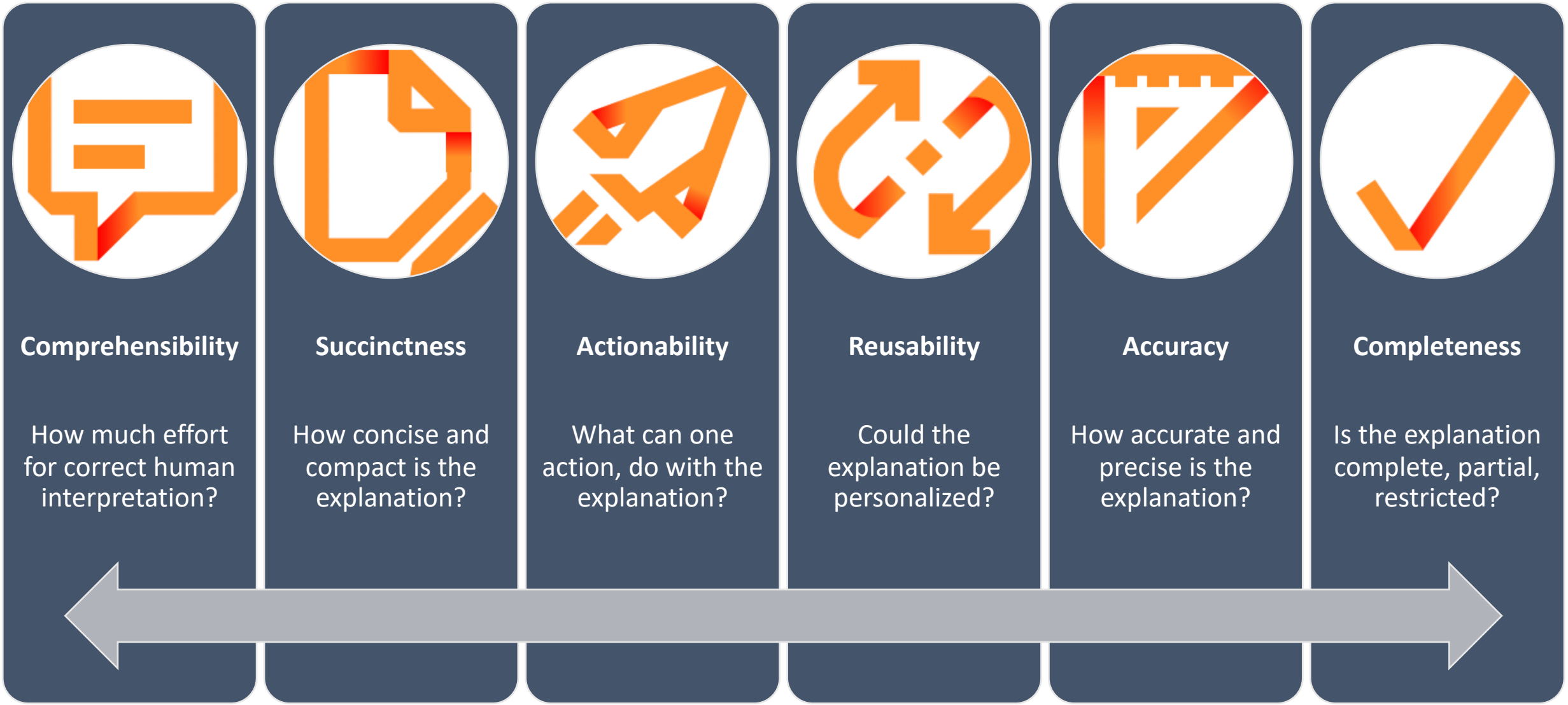
Perturbation-based Evaluation for Feature Attribution-based Approaches

Perturb top-k features by attribution and observe change in prediction

- Higher the change, better the method
- Perturbation may amount to replacing the feature with a random value
- Samek et al. formalize this using a metric: **Area over perturbation curve**
  - Plot the prediction for input with top-k features perturbed as a function of k
  - Take the area over this curve



XAI: One Objective, Many Metrics



Source: Accenture Point of View. Understanding Machines: Explainable AI. Freddy Lecue, Dadong Wan

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]

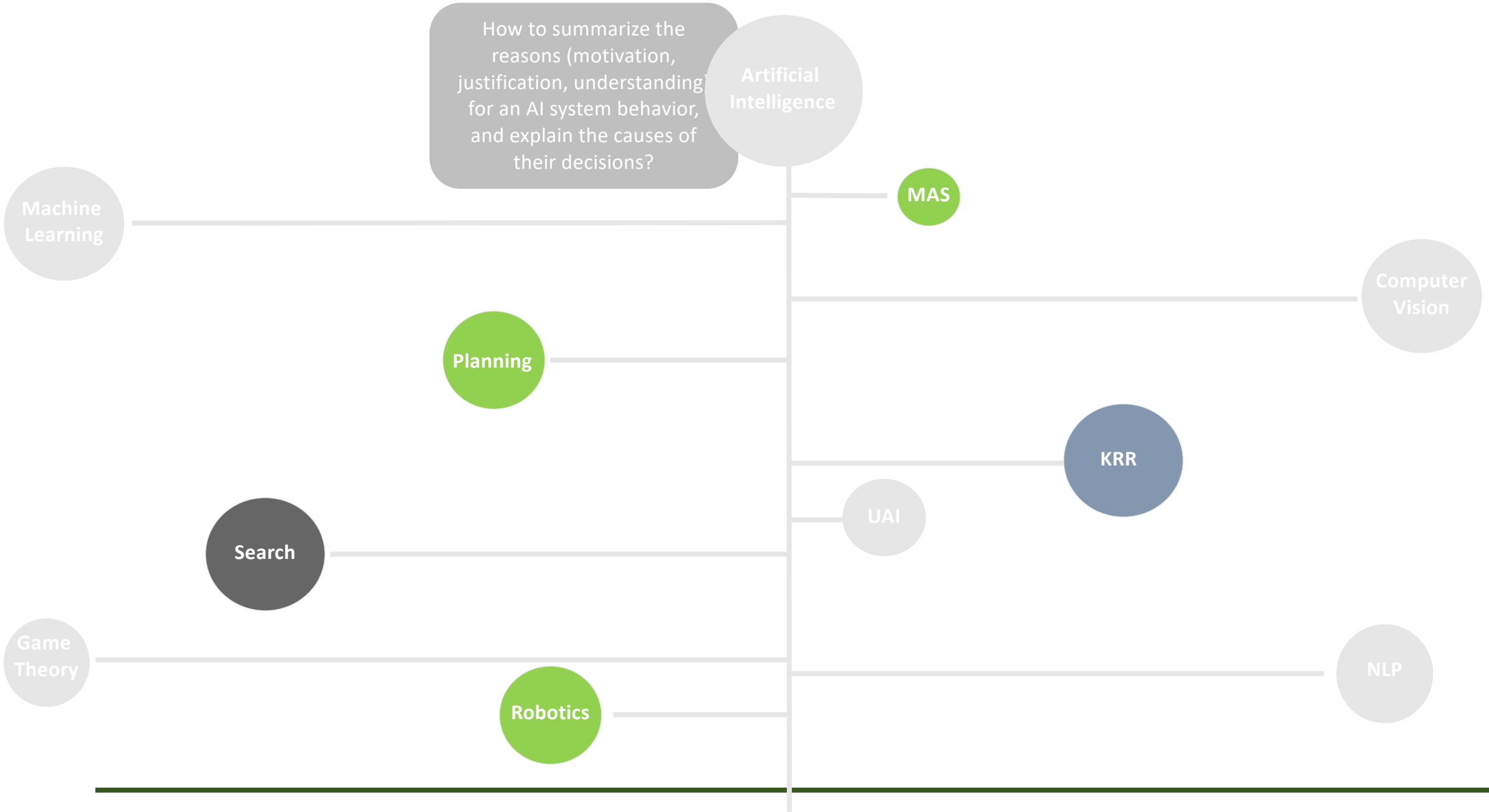
tl;dr

- 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

XAI in AI

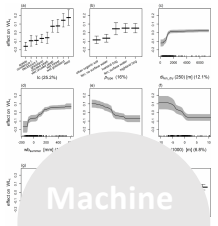
XAI: One Objective, Many 'AI's, Many Definitions, Many Approaches

XAI: One Objective, Many 'AI's, Many Definitions, Many Approaches

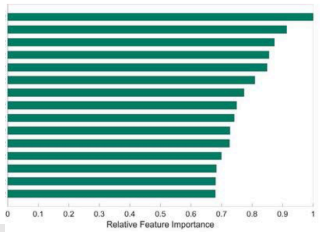


XAI: One Objective, Many 'AI's, Many Definitions, Many Approaches

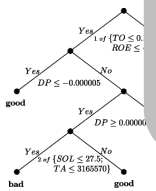
Dependency Plot



Feature Importance



Surrogate Model



Machine Learning

Which features are responsible of classification?

How to summarize the reasons (motivation, justification, understanding) for an AI system behavior, and explain the causes of their decisions?

Artificial Intelligence

MAS

Computer Vision

Planning

KRR

UAI

Search

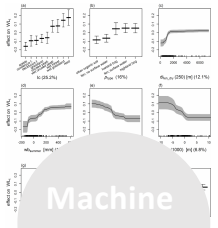
Game Theory

NLP

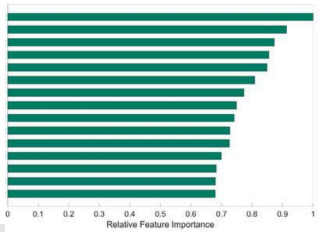
Robotics

XAI: One Objective, Many 'AI's, Many Definitions, Many Approaches

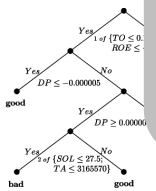
Dependency Plot



Feature Importance



Surrogate Model



How to summarize the reasons (motivation, justification, understanding) for an AI system behavior, and explain the causes of their decisions?

Artificial Intelligence

Machine Learning

Which features are responsible of classification?

Planning

Search

Game Theory

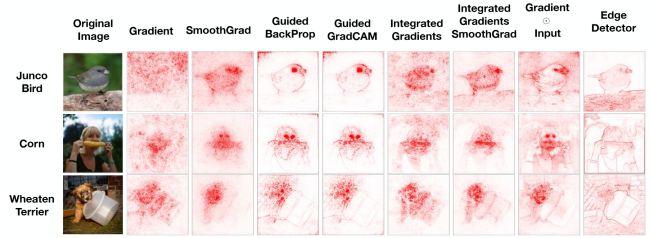
Robotics

MAS

KRR

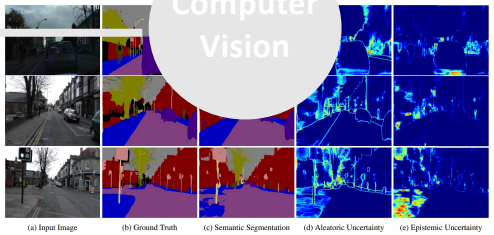
UAI

Saliency Map



Which complex features are responsible of classification?

Computer Vision

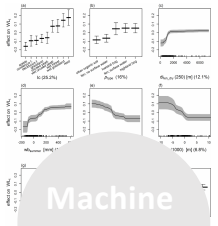


Uncertainty Map

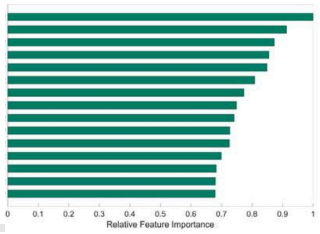
NLP

XAI: One Objective, Many ‘AI’s, Many Definitions, Many Approaches

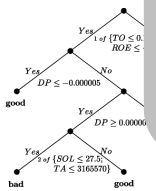
Dependency Plot



Feature Importance



Surrogate Model



Machine Learning

Which features are responsible of classification?

How to summarize the reasons (motivation, justification, understanding) for an AI system behavior, and explain the causes of their decisions?

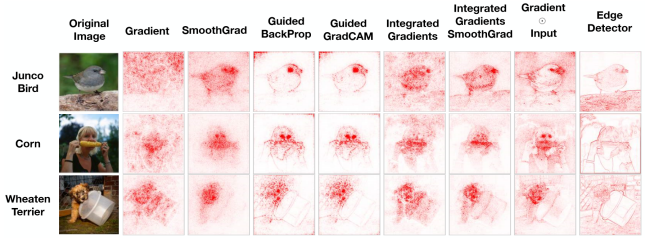
Artificial Intelligence

Strategy Summarization

MAS

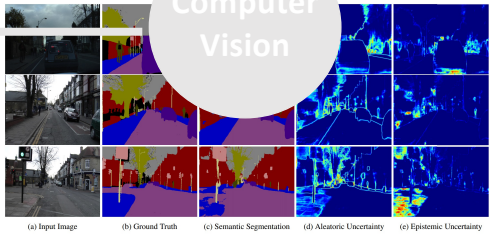
- Which agent strategy & plan ?
- Which player contributes most?
- Why such a conversational flow?

Saliency Map



Which complex features are responsible of classification?

Computer Vision



Uncertainty Map

Planning

KRR

UAI

Search

Game Theory

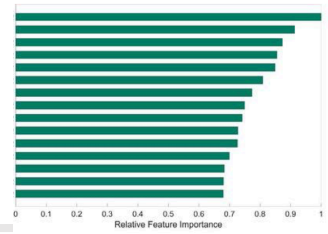
Robotics

NLP

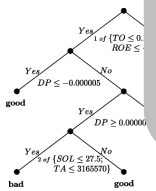
XAI: One Objective, Many 'AI's, Many Definitions, Many Approaches

Dependency Plot

Feature Importance



Surrogate Model



Machine Learning

Which features are responsible of classification?

How to summarize the reasons (motivation, justification, understanding) for an AI system behavior, and explain the causes of their decisions?

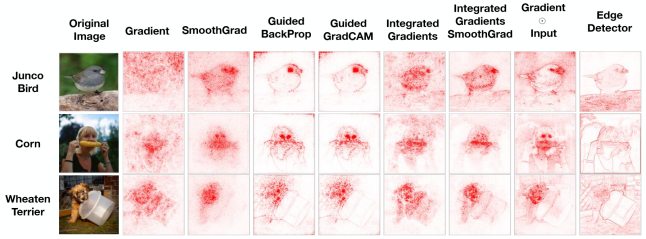
Artificial Intelligence

Strategy Summarization

MAS

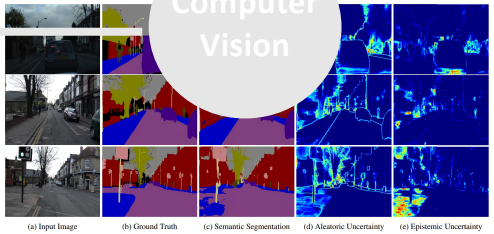
- Which agent strategy & plan ?
- Which player contributes most?
- Why such a conversational flow?

Saliency Map



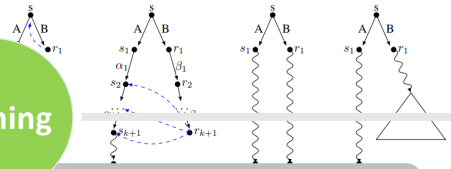
Which complex features are responsible of classification?

Computer Vision



Uncertainty Map

Plan Refinement



Planning

Which actions are responsible of a plan?

KRR

UAI

Search

Game Theory

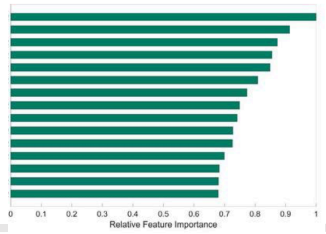
Robotics

NLP

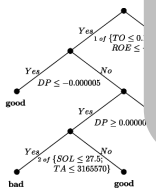
XAI: One Objective, Many 'AI's, Many Definitions, Many Approaches

Dependency Plot

Feature Importance



Surrogate Model



Machine Learning

Which features are responsible of classification?

How to summarize the reasons (motivation, justification, understanding) for an AI system behavior, and explain the causes of their decisions?

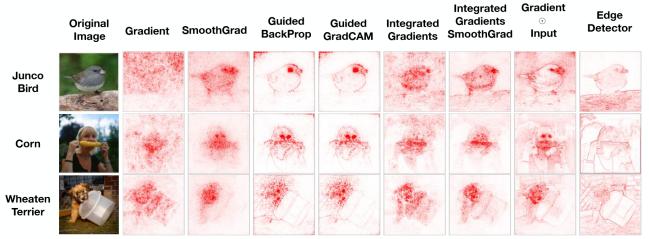
Artificial Intelligence

Strategy Summarization

MAS

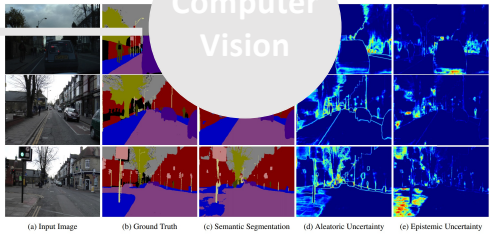
- Which agent strategy & plan ?
- Which player contributes most?
- Why such a conversational flow?

Saliency Map



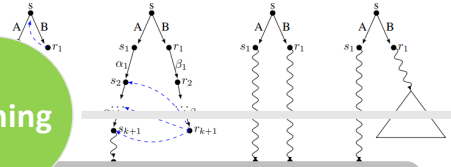
Which complex features are responsible of classification?

Computer Vision



Uncertainty Map

Plan Refinement



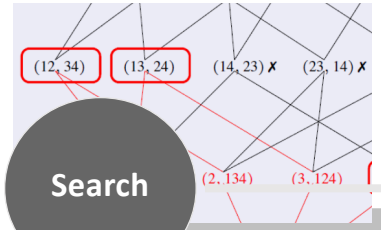
Planning

Which actions are responsible of a plan?

KRR

UAI

Conflicts Resolution



Search

Which constraints can be relaxed?

Game Theory

Robotics

NLP

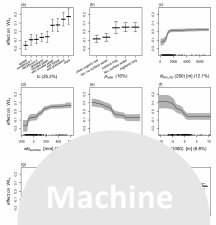
XAI: One Objective, Many 'AI's, Many Definitions, Many Approaches

How to summarize the reasons (motivation, justification, understanding) for an AI system behavior, and explain the causes of their decisions?

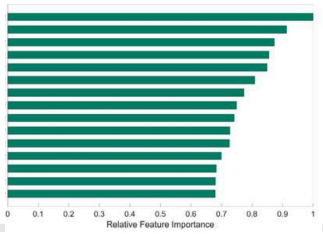
Artificial Intelligence

Machine Learning

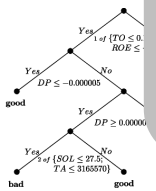
Dependency Plot



Feature Importance



Surrogate Model



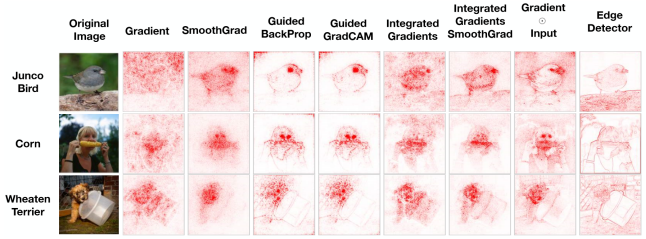
Which features are responsible of classification?

MAS

Strategy Summarization

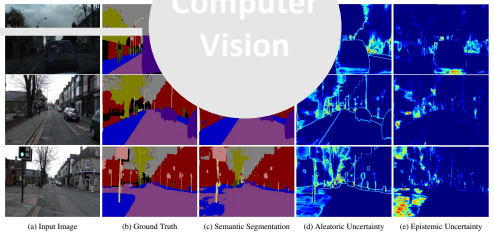
- Which agent strategy & plan ?
- Which player contributes most?
- Why such a conversational flow?

Saliency Map



Which complex features are responsible of classification?

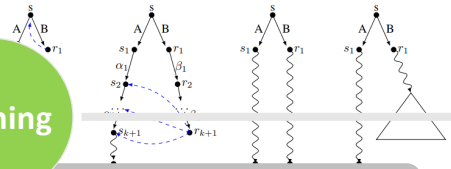
Computer Vision



Uncertainty Map

Planning

Plan Refinement



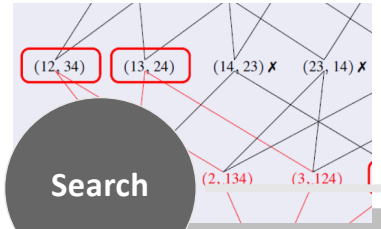
Which actions are responsible of a plan?

KRR

UAI

Conflicts Resolution

Search



Which constraints can be relaxed?

Game Theory

Which combination of features is optimal?

Robotics

NLP

Shapely Values

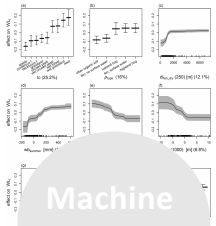
XAI: One Objective, Many 'AI's, Many Definitions, Many Approaches

How to summarize the reasons (motivation, justification, understanding) for an AI system behavior, and explain the causes of their decisions?

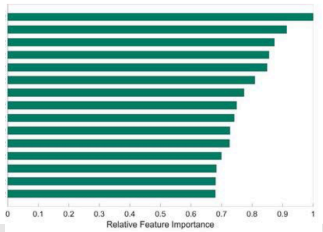
Artificial Intelligence

Machine Learning

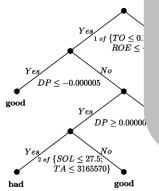
Dependency Plot



Feature Importance



Surrogate Model



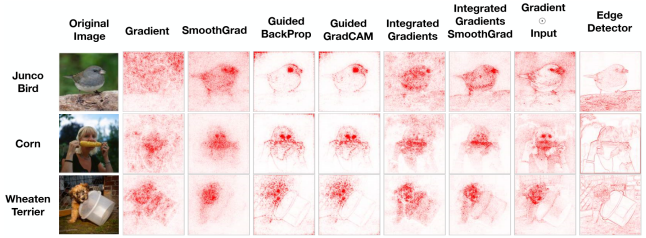
Which features are responsible of classification?

MAS

Strategy Summarization

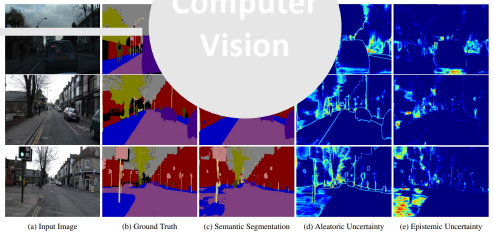
- Which agent strategy & plan ?
- Which player contributes most?
- Why such a conversational flow?

Saliency Map



Which complex features are responsible of classification?

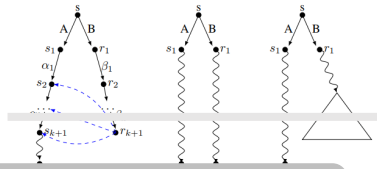
Computer Vision



Uncertainty Map

Planning

Plan Refinement



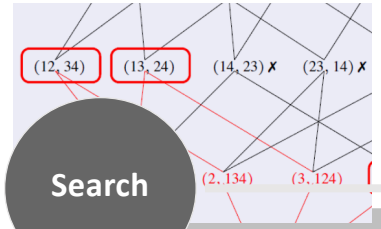
Which actions are responsible of a plan?

KRR

UAI

Conflicts Resolution

Search



Which constraints can be relaxed?

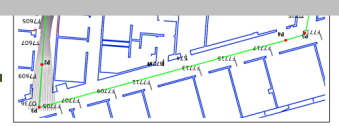
Game Theory

Which combination of features is optimal?

Shapely Values

Robotics

Which decisions, combination of multimodal decisions lead to an action?



Narrative-based

NLP

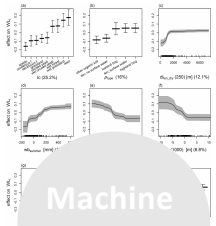
XAI: One Objective, Many 'AI's, Many Definitions, Many Approaches

How to summarize the reasons (motivation, justification, understanding) for an AI system behavior, and explain the causes of their decisions?

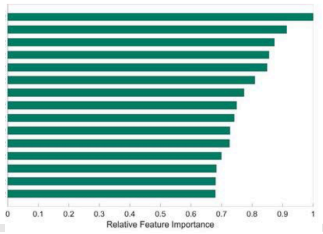
Artificial Intelligence

Machine Learning

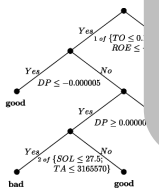
Dependency Plot



Feature Importance



Surrogate Model



Which features are responsible of classification?

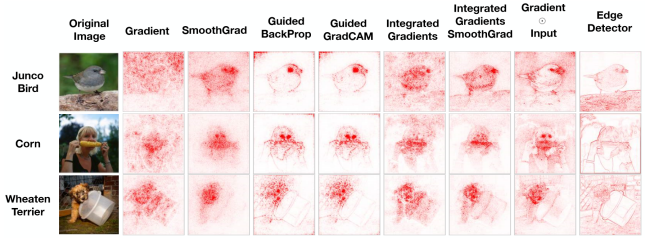
MAS

Strategy Summarization

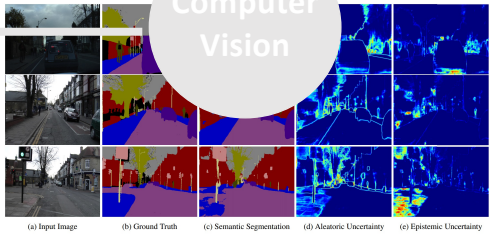
- Which agent strategy & plan ?
- Which player contributes most?
- Why such a conversational flow?

Computer Vision

Saliency Map



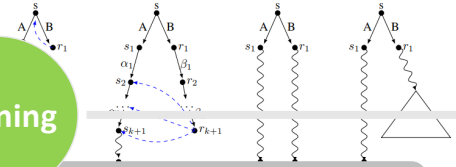
Which complex features are responsible of classification?



Uncertainty Map

Planning

Plan Refinement



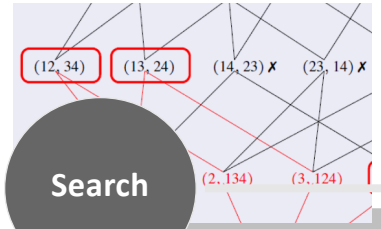
Which actions are responsible of a plan?

KRR

UAI

Conflicts Resolution

Search



Which constraints can be relaxed?

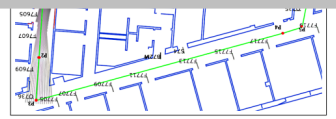
Game Theory

Which combination of features is optimal?

Shapely Values

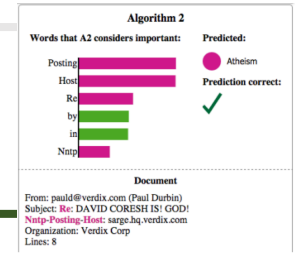
Robotics

Which decisions, combination of multimodal decisions lead to an action?



Narrative-based

Machine Learning based



NLP

Which entity is responsible for classification?

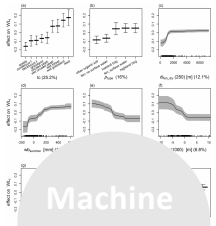
XAI: One Objective, Many 'AI's, Many Definitions, Many Approaches

How to summarize the reasons (motivation, justification, understanding) for an AI system behavior, and explain the causes of their decisions?

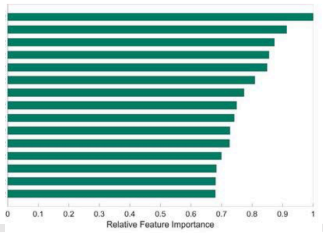
Artificial Intelligence

Machine Learning

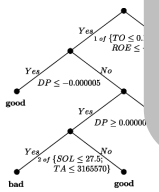
Dependency Plot



Feature Importance



Surrogate Model



Which features are responsible of classification?

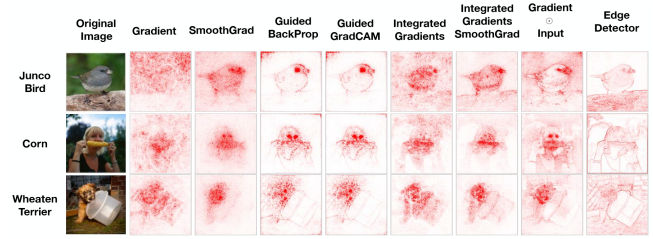
MAS

Strategy Summarization

- Which agent strategy & plan ?
- Which player contributes most?
- Why such a conversational flow?

Computer Vision

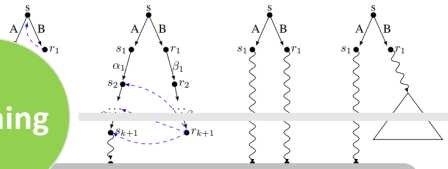
Saliency Map



Which complex features are responsible of classification?

Planning

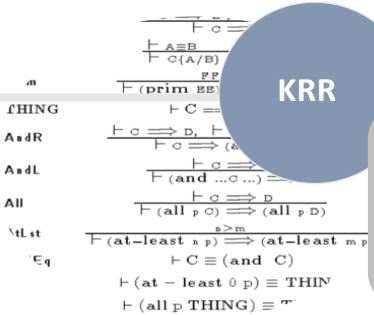
Plan Refinement



Which actions are responsible of a plan?

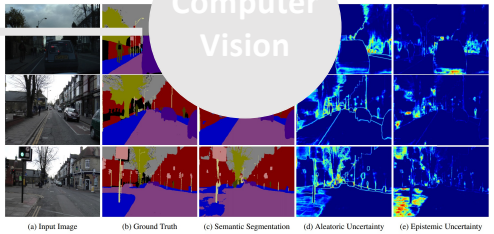
KRR

Diagnosis



Abduction

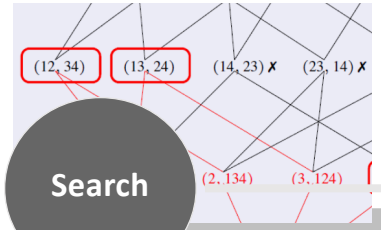
Uncertainty Map



- Which axiom is responsible of inference (e.g., classification)?
- Abduction/Diagnostic: Find the **right** root causes (abduction)?

Search

Conflicts Resolution



Which constraints can be relaxed?

UAI

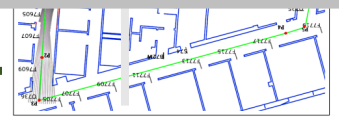
Game Theory

Which combination of features is optimal?

Shapely Values

Robotics

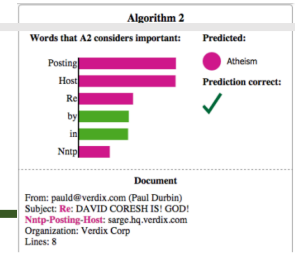
Which decisions, combination of multimodal decisions lead to an action?



Narrative-based

NLP

Machine Learning based



Which entity is responsible for classification?

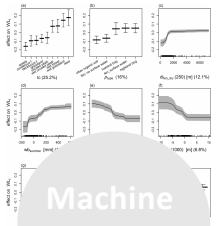
XAI: One Objective, Many 'AI's, Many Definitions, Many Approaches

How to summarize the reasons (motivation, justification, understanding) for an AI system behavior, and explain the causes of their decisions?

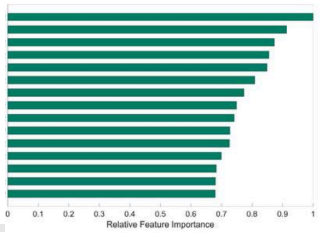
Artificial Intelligence

Machine Learning

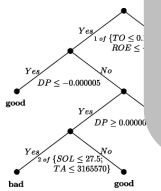
Dependency Plot



Feature Importance



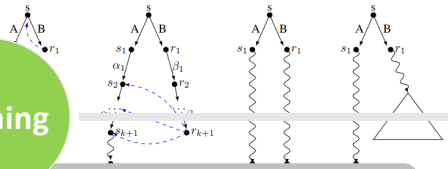
Surrogate Model



Which features are responsible of classification?

Planning

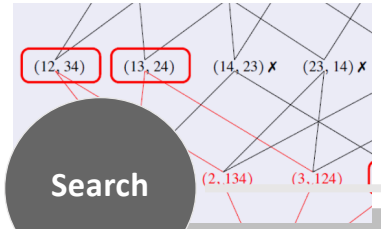
Plan Refinement



Which actions are responsible of a plan?

Search

Conflicts Resolution



Which constraints can be relaxed?

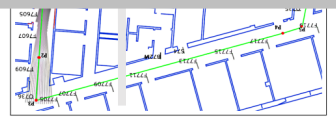
Game Theory

Which combination of features is optimal?

Shapely Values

Robotics

Which decisions, combination of multimodal decisions lead to an action?



Narrative-based

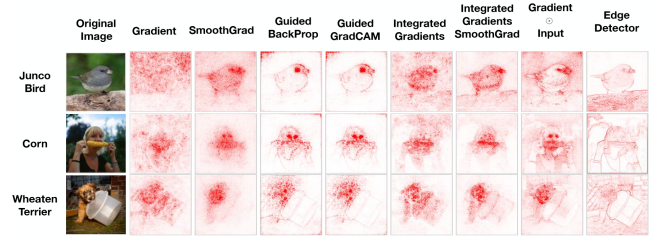
MAS

Strategy Summarization

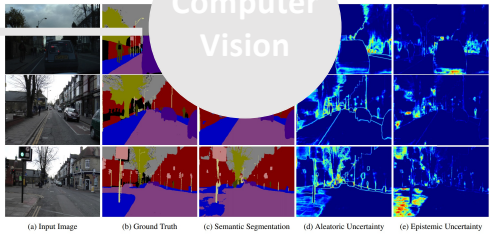
- Which agent strategy & plan ?
- Which player contributes most?
- Why such a conversational flow?

Computer Vision

Saliency Map



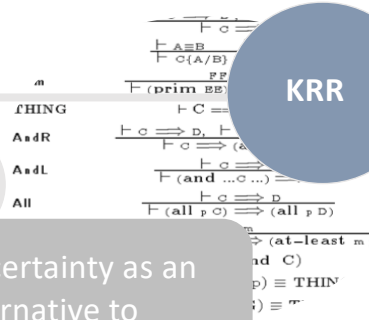
Which complex features are responsible of classification?



Uncertainty Map

KRR

Diagnosis



Abduction

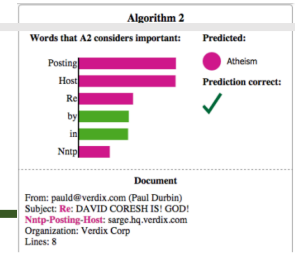
- Which axiom is responsible of inference (e.g., classification)?
- Abduction/Diagnostic: Find the **right** root causes (abduction)?

UAI

Uncertainty as an alternative to explanation

Machine Learning based

NLP



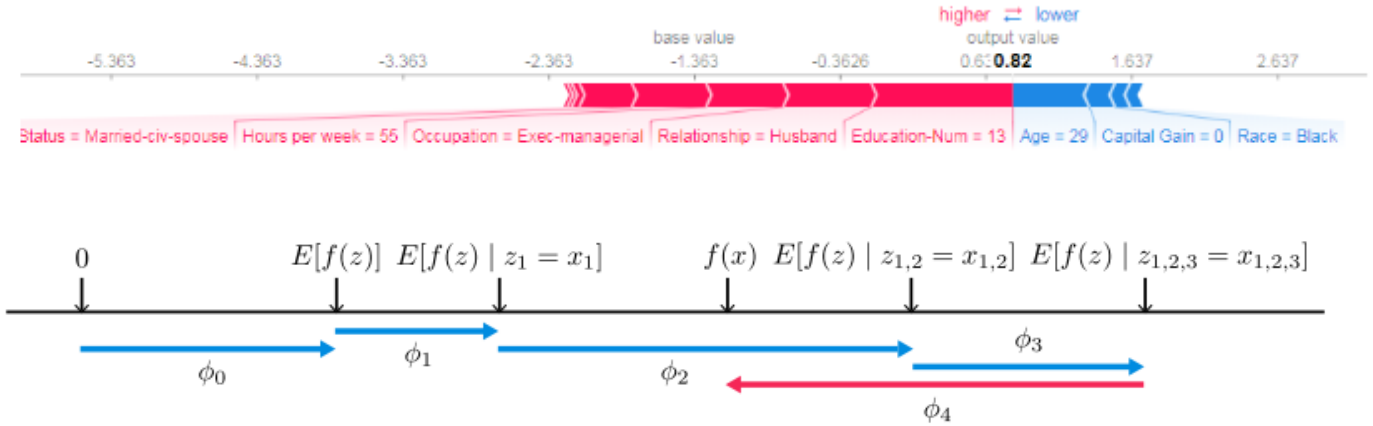
Which entity is responsible for classification?

Deep Dive on XAI in AI

(except Machine Learning)

Overview of explanation in different AI fields (1)

- Game Theory

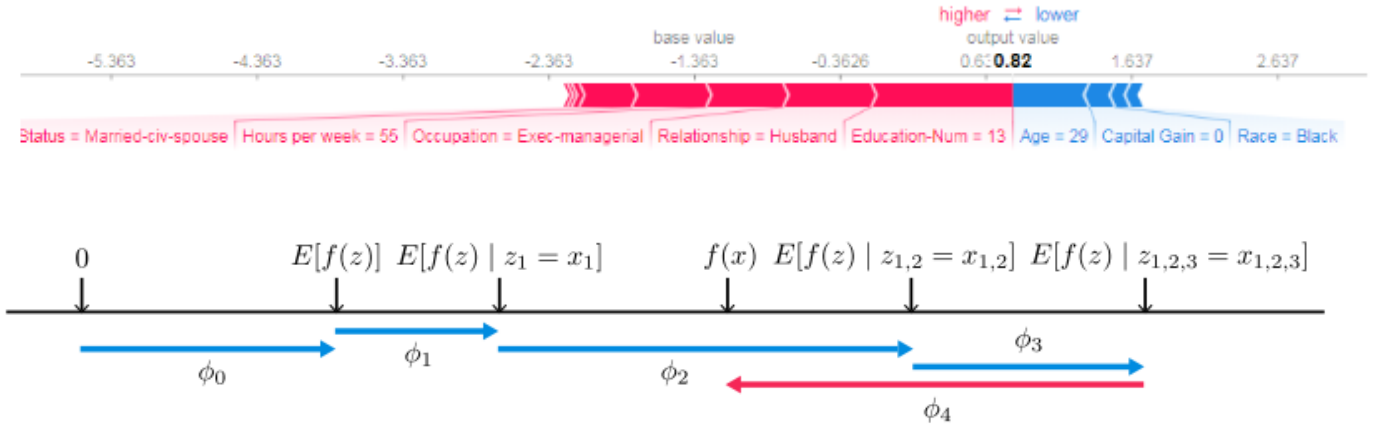


Shapley Additive Explanation

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

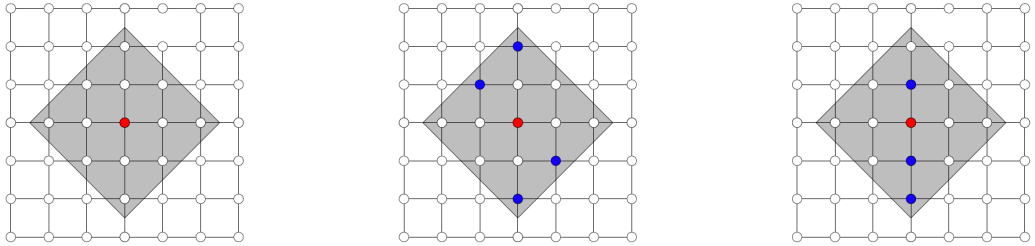
Overview of explanation in different AI fields (1)

- Game Theory



Shapley Additive Explanation

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

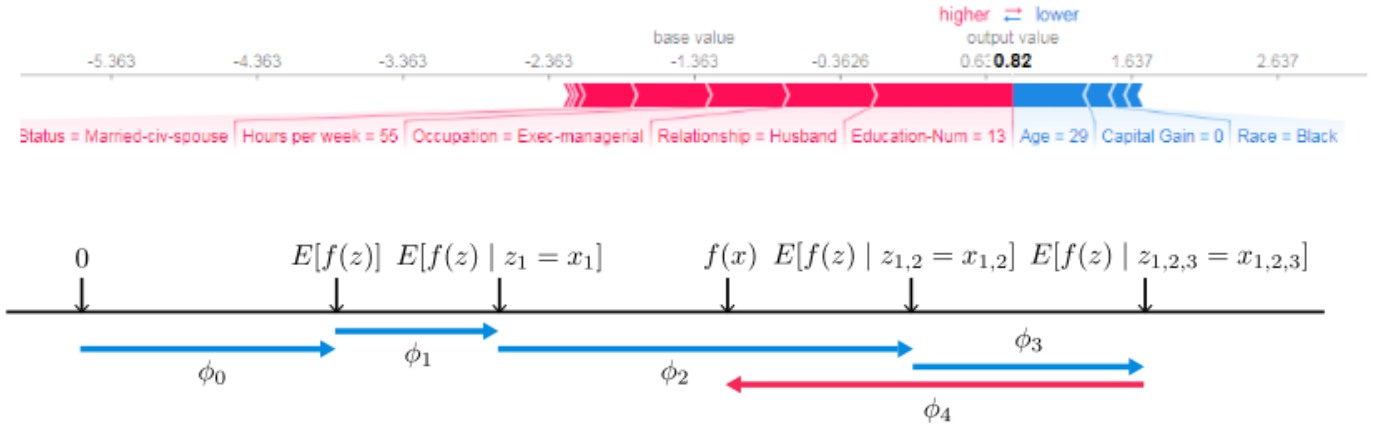


L-Shapley and C-Shapley (with graph structure)

Jianbo Chen, Le Song, Martin J. Wainwright, Michael I. Jordan: L-Shapley and C-Shapley: Efficient Model Interpretation for Structured Data. ICLR 2019

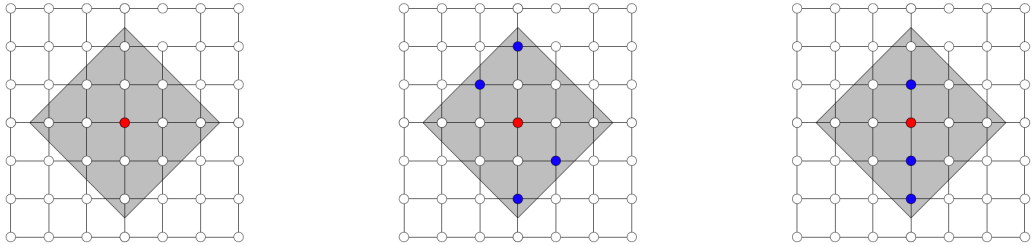
Overview of explanation in different AI fields (1)

- Game Theory



Shapley Additive Explanation

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



L-Shapley and C-Shapley (with graph structure)

Jianbo Chen, Le Song, Martin J. Wainwright, Michael I. Jordan: L-Shapley and C-Shapley: Efficient Model Interpretation for Structured Data. ICLR 2019

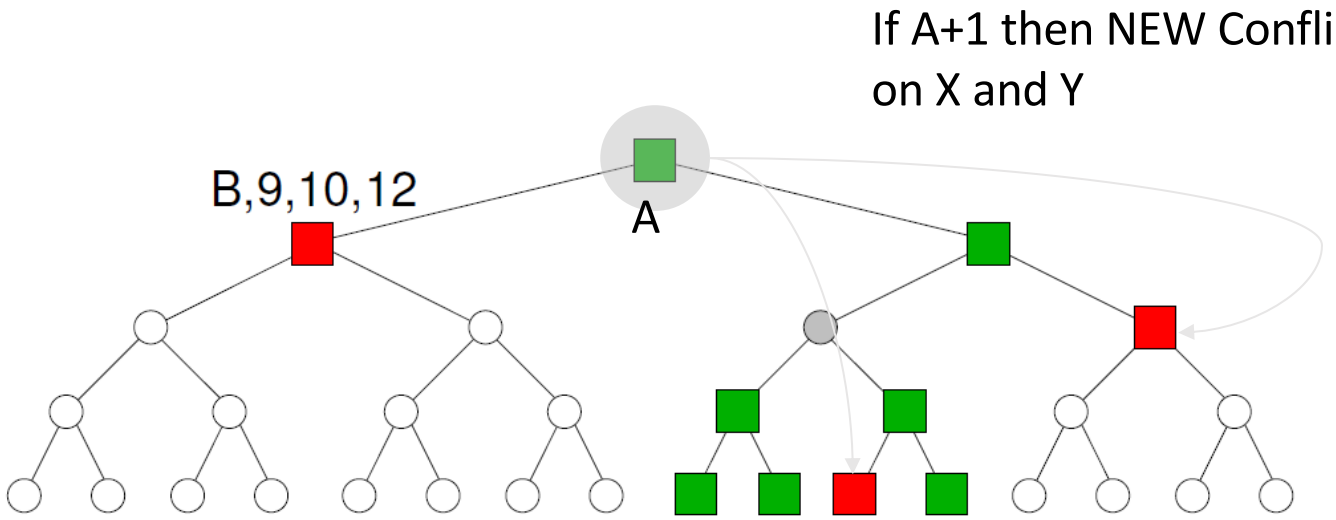
~ instancewise feature importance (causal influence)

Erik Štrumbelj and Igor Kononenko. An efficient explanation of individual classifications using game theory. Journal of Machine Learning Research, 11:1–18, 2010.

Anupam Datta, Shayak Sen, and Yair Zick. Algorithmic transparency via quantitative input influence: Theory and experiments with learning systems. In Security and Privacy (SP), 2016 IEEE Symposium on, pp. 598–617. IEEE, 2016.

Overview of explanation in different AI fields (2)

- Search and Constraint Satisfaction



Conflicts resolution

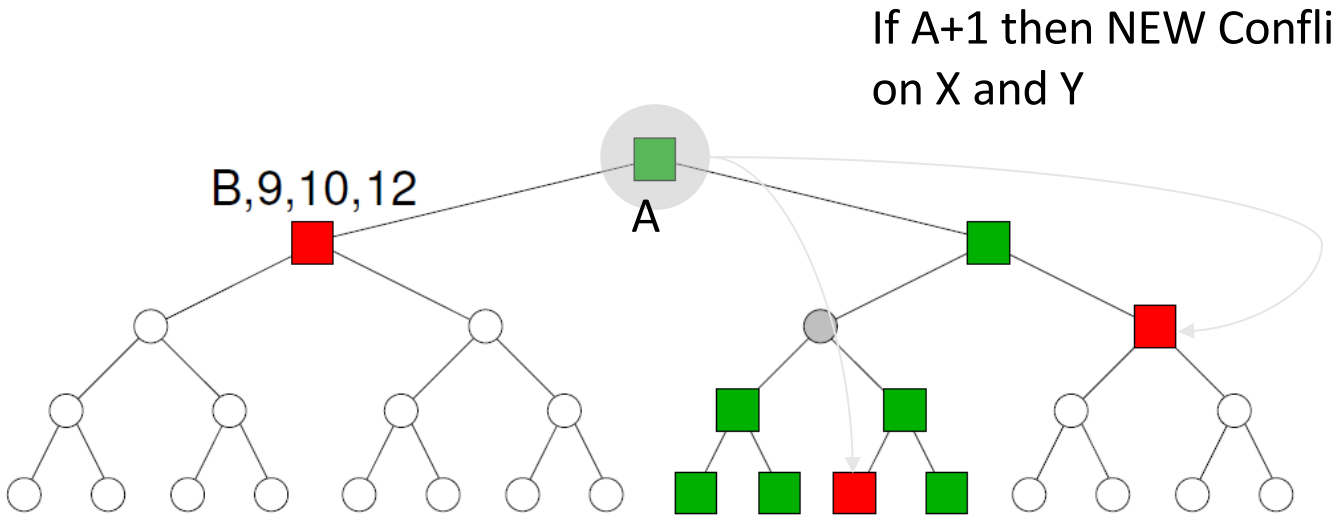
Barry O'Sullivan, Alexandre Papadopoulos, Boi Faltings, Pearl Pu: Representative Explanations for Over-Constrained Problems. AAAI 2007: 323-328

Robustness Computation

Hebrard, E., Hnich, B., & Walsh, T. (2004, July). Robust solutions for constraint satisfaction and optimization. In ECAI (Vol. 16, p. 186).

Overview of explanation in different AI fields (2)

- Search and Constraint Satisfaction

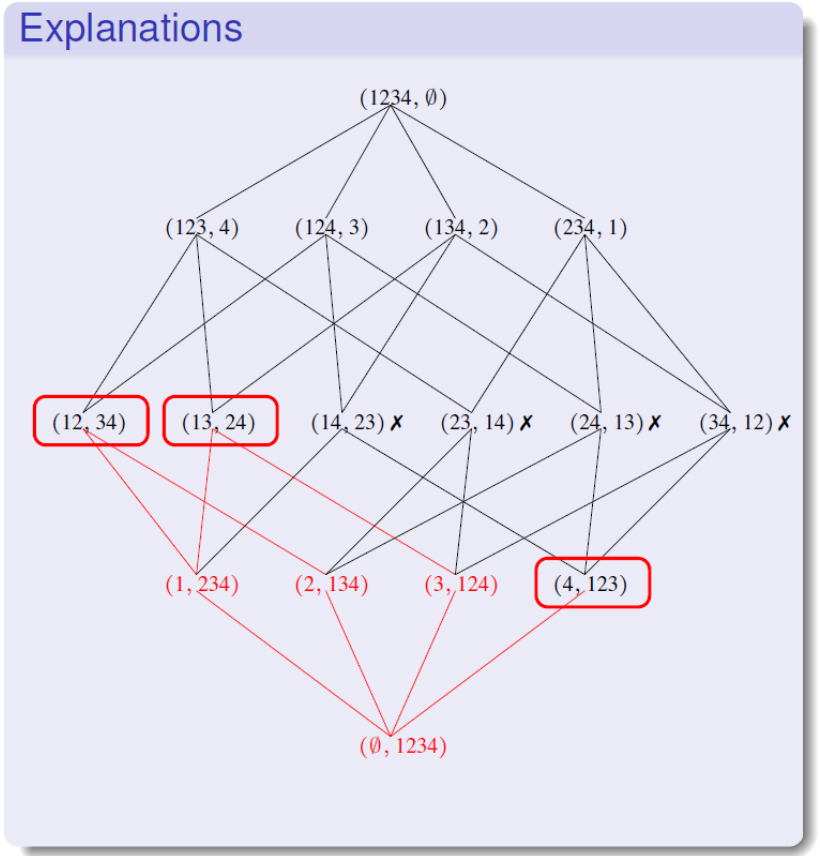


Conflicts resolution

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

Robustness Computation

Hebrard, E., Hnich, B., & Walsh, T. (2004, July). Robust solutions for constraint satisfaction and optimization. In ECAI (Vol. 16, p. 186).



Constraints relaxation

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

Overview of explanation in different AI fields (3)

- Knowledge Representation and Reasoning

Ref	$\vdash C \Longrightarrow C$
Trans	$\dfrac{\vdash C \Longrightarrow D,\ \vdash D \Longrightarrow E}{\vdash C \Longrightarrow E}$
Eq	$\dfrac{\vdash A \equiv B \qquad \vdash C \Longrightarrow D}{\vdash C\{A/B\} \Longrightarrow D\{A/B\}}$
Prim	$\dfrac{FF \subset EE}{\vdash (\text{prim } EE) \Longrightarrow (\text{prim } FF)}$
THING	$\vdash C \Longrightarrow \text{THING}$
AndR	$\dfrac{\vdash C \Longrightarrow D,\ \vdash C \Longrightarrow (\text{and } EE)}{\vdash C \Longrightarrow (\text{and } D\ EE)}$
AndL	$\dfrac{\vdash C \Longrightarrow E}{\vdash (\text{and } \ldots C \ldots) \Longrightarrow E}$
All	$\dfrac{\vdash C \Longrightarrow D}{\vdash (\text{all } p\ C) \Longrightarrow (\text{all } p\ D)}$
AtLst	$\dfrac{n > m}{\vdash (\text{at-least } n\ p) \Longrightarrow (\text{at-least } m\ p)}$
AndEq	$\vdash C \equiv (\text{and } C)$
AtLs0	$\vdash (\text{at} - \text{least } 0\ p) \equiv \text{THING}$
All-thing	$\vdash (\text{all } p\ \text{THING}) \equiv \text{THING}$
All-and	$\vdash (\text{and } (\text{all } p\ C)(\text{all } p\ D) \ldots) \equiv (\text{and } (\text{all } p\ (\text{and } C\ D)) \ldots)$

1.	(at-least 3 grape) $\Longrightarrow$ (at-least 2 grape)	AtLst
2.	(and (at-least 3 grape) (prim GOOD WINE)) $\Longrightarrow$ (at-least 2 grape)	AndL,1
3.	(prim GOOD WINE) $\Longrightarrow$ (prim WINE)	Prim
4.	(and (at-least 3 grape) (prim GOOD WINE)) $\Longrightarrow$ (prim WINE)	AndL,3
5.	A $\equiv$ (and (at-least 3 grape) (prim GOOD WINE))	Told
6.	A $\Longrightarrow$ (prim WINE)	Eq,4,5
7.	(prim WINE) $\equiv$ (and (prim WINE))	AndEq
8.	A $\Longrightarrow$ (and (prim WINE))	Eq,7,6
9.	A $\Longrightarrow$ (at-least 2 grape)	Eq,5,2
10.	A $\Longrightarrow$ (and (at-least 2 grape) (prim WINE))	AndR,9,8

A $\equiv$ (and (at-least 3 grape) (prim GOOD WINE))

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

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

Overview of explanation in different AI fields (3)

- Knowledge Representation and Reasoning

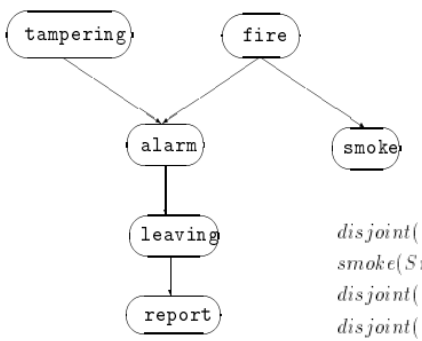
Ref	$\vdash C \Longrightarrow C$
Trans	$\frac{\vdash C \Longrightarrow D,\ \vdash D \Longrightarrow E}{\vdash C \Longrightarrow E}$
Eq	$\frac{\vdash A \equiv B \qquad \vdash C \Longrightarrow D}{\vdash C\{A/B\} \Longrightarrow D\{A/B\}}$
Prim	$\frac{FF \subset EE}{\vdash (\text{prim } EE) \Longrightarrow (\text{prim } FF)}$
THING	$\vdash C \Longrightarrow \text{THING}$
AndR	$\frac{\vdash C \Longrightarrow D,\ \vdash C \Longrightarrow (\text{and } EE)}{\vdash C \Longrightarrow (\text{and } D\ EE)}$
AndL	$\frac{\vdash C \Longrightarrow E}{\vdash (\text{and } \ldots C \ldots) \Longrightarrow E}$
All	$\frac{\vdash C \Longrightarrow D}{\vdash (\text{all } p\ C) \Longrightarrow (\text{all } p\ D)}$
AtLst	$\frac{n > m}{\vdash (\text{at-least } n\ p) \Longrightarrow (\text{at-least } m\ p)}$
AndEq	$\vdash C \equiv (\text{and } C)$
AtLs0	$\vdash (\text{at} - \text{least } 0\ p) \equiv \text{THING}$
All-thing	$\vdash (\text{all } p\ \text{THING}) \equiv \text{THING}$
All-and	$\vdash (\text{and } (\text{all } p\ C)(\text{all } p\ D) \ldots) \equiv (\text{and } (\text{all } p\ (\text{and } C\ D)) \ldots)$

1. (at-least 3 grape) $\Longrightarrow$ (at-least 2 grape)		AtLst
2. (and (at-least 3 grape) (prim GOOD WINE)) $\Longrightarrow$ (at-least 2 grape)		AndL,1
3. (prim GOOD WINE) $\Longrightarrow$ (prim WINE)		Prim
4. (and (at-least 3 grape) (prim GOOD WINE)) $\Longrightarrow$ (prim WINE)		AndL,3
5. A $\equiv$ (and (at-least 3 grape) (prim GOOD WINE))		Told
6. A $\Longrightarrow$ (prim WINE)		Eq,4,5
7. (prim WINE) $\equiv$ (and (prim WINE))		AndEq
8. A $\Longrightarrow$ (and (prim WINE))		Eq,7,6
9. A $\Longrightarrow$ (at-least 2 grape)		Eq,5,2
10. A $\Longrightarrow$ (and (at-least 2 grape) (prim WINE))		AndR,9,8

A $\equiv$ (and (at-least 3 grape) (prim GOOD WINE))

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

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



$$P(alarm|fire \wedge \neg tampering) = 0.99$$
$$P(alarm|\neg fire \wedge tampering) = 0.85$$
$$P(alarm|\neg fire \wedge \neg tampering) = 0.0001$$
$$P(leaving|alarm) = 0.88$$
$$P(leaving|\neg alarm) = 0.001$$
$$P(\text{report}|leaving) = 0.75$$
$$P(\text{report}|\neg leaving) = 0.01$$

$disjoint([fire(yes) : 0.01, fire(no) : 0.99]).$
$smoke(Sm) \leftarrow fire(Fi) \wedge c\_smoke(Sm, Fi).$
$disjoint([c\_smoke(yes, yes) : 0.9, c\_smoke(no, yes) : 0.1]).$
$disjoint([c\_smoke(yes, no) : 0.01, c\_smoke(no, no) : 0.99]).$

Abduction Reasoning (in Bayesian Network)

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

Overview of explanation in different AI fields (3)

- Knowledge Representation and Reasoning

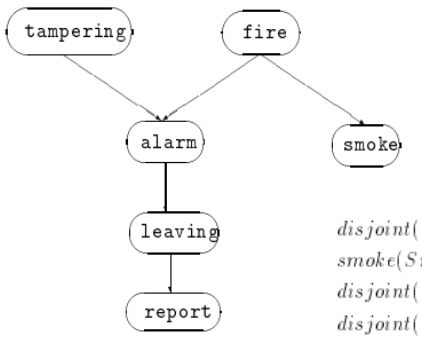
Ref	$\vdash C \Longrightarrow C$
Trans	$\dfrac{\vdash C \Longrightarrow D,\ \vdash D \Longrightarrow E}{\vdash C \Longrightarrow E}$
Eq	$\dfrac{\vdash A \equiv B \qquad \vdash C \Longrightarrow D}{\vdash C\{A/B\} \Longrightarrow D\{A/B\}}$
Prim	$\dfrac{FF \subset EE}{\vdash (\text{prim } EE) \Longrightarrow (\text{prim } FF)}$
THING	$\vdash C \Longrightarrow \text{THING}$
AndR	$\dfrac{\vdash C \Longrightarrow D,\ \vdash C \Longrightarrow (\text{and } EE)}{\vdash C \Longrightarrow (\text{and } D\ EE)}$
AndL	$\dfrac{\vdash C \Longrightarrow E}{\vdash (\text{and } \ldots C \ldots) \Longrightarrow E}$
All	$\dfrac{\vdash C \Longrightarrow D}{\vdash (\text{all } p\ C) \Longrightarrow (\text{all } p\ D)}$
AtLst	$\dfrac{n > m}{\vdash (\text{at-least } n\ p) \Longrightarrow (\text{at-least } m\ p)}$
AndEq	$\vdash C \equiv (\text{and } C)$
AtLs0	$\vdash (\text{at-least } 0\ p) \equiv \text{THING}$
All-thing	$\vdash (\text{all } p\ \text{THING}) \equiv \text{THING}$
All-and	$\vdash (\text{and } (\text{all } p\ C)(\text{all } p\ D) \ldots) \equiv (\text{and } (\text{all } p\ (\text{and } C\ D)) \ldots)$

1. (at-least 3 grape) $\Longrightarrow$ (at-least 2 grape)		AtLst
2. (and (at-least 3 grape) (prim GOOD WINE)) $\Longrightarrow$ (at-least 2 grape)		AndL,1
3. (prim GOOD WINE) $\Longrightarrow$ (prim WINE)		Prim
4. (and (at-least 3 grape) (prim GOOD WINE)) $\Longrightarrow$ (prim WINE)		AndL,3
5. A $\equiv$ (and (at-least 3 grape) (prim GOOD WINE))		Told
6. A $\Longrightarrow$ (prim WINE)		Eq,4,5
7. (prim WINE) $\equiv$ (and (prim WINE))		AndEq
8. A $\Longrightarrow$ (and (prim WINE))		Eq,7,6
9. A $\Longrightarrow$ (at-least 2 grape)		Eq,5,2
10. A $\Longrightarrow$ (and (at-least 2 grape) (prim WINE))		AndR,9,8

A $\equiv$ (and (at-least 3 grape) (prim GOOD WINE))

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

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

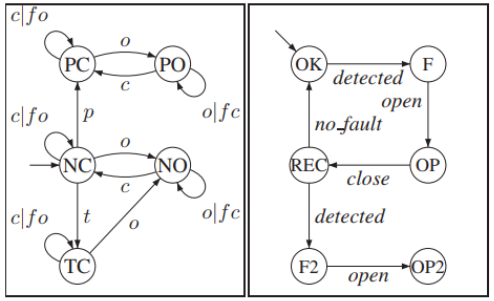


$$P(alarm|fire \wedge \neg tampering) = 0.99$$
$$P(alarm|\neg fire \wedge tampering) = 0.85$$
$$P(alarm|\neg fire \wedge \neg tampering) = 0.0001$$
$$P(leaving|alarm) = 0.88$$
$$P(leaving|\neg alarm) = 0.001$$
$$P(\text{report}|leaving) = 0.75$$
$$P(\text{report}|\neg leaving) = 0.01$$

$$disjoint([fire(yes) : 0.01, fire(no) : 0.99]).$$
$$smoke(Sm) \leftarrow fire(Fi) \wedge c\_smoke(Sm, Fi).$$
$$disjoint([c\_smoke(yes, yes) : 0.9, c\_smoke(no, yes) : 0.1]).$$
$$disjoint([c\_smoke(yes, no) : 0.01, c\_smoke(no, no) : 0.99]).$$

Abduction Reasoning (in Bayesian Network)

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

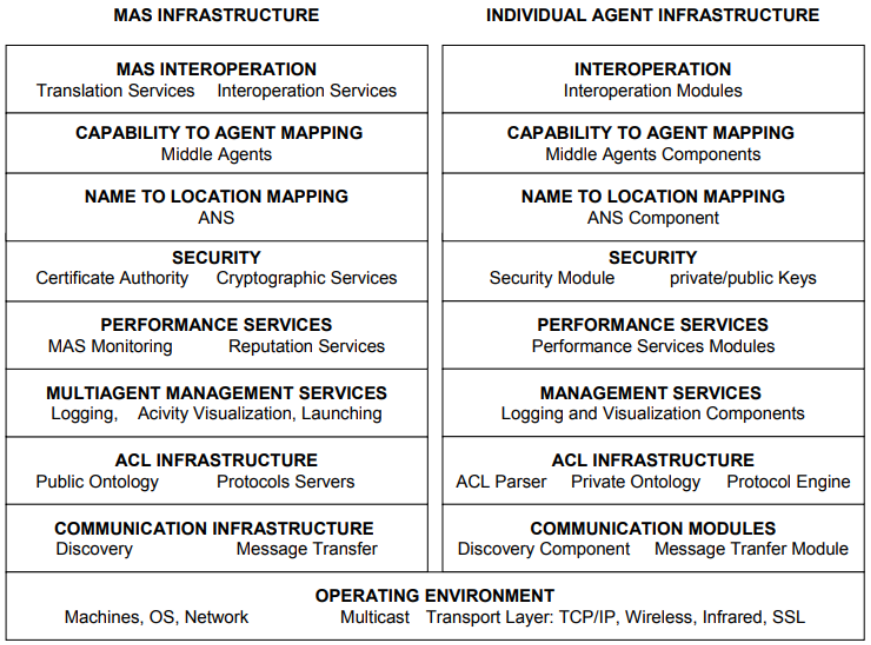


Diagnosis Inference

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 (4)

- Multi-agent Systems



Explanation of Agent Conflicts & 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)

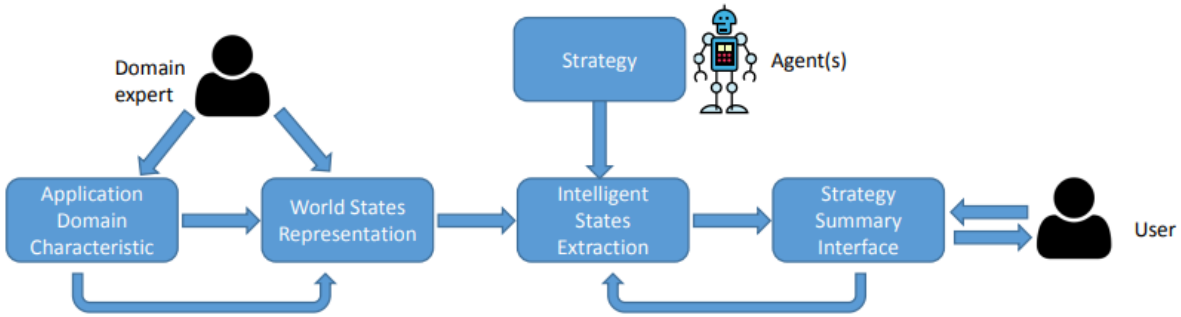
Overview of explanation in different AI fields (4)

- 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, Acivity 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 Tranfer Module
OPERATING ENVIRONMENT Machines, OS, Network Multicast Transport Layer: TCP/IP, Wireless, Infrared, SSL	

Explanation of Agent Conflicts & 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

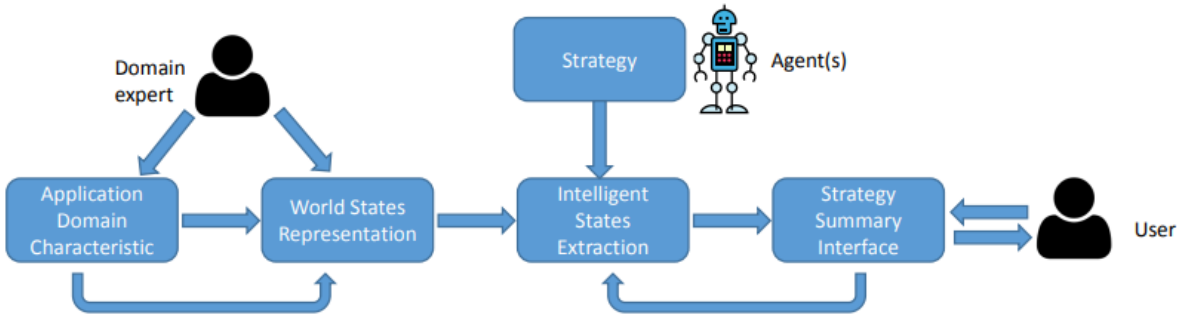
Overview of explanation in different AI fields (4)

- 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, Acivity 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 Tranfer Module
OPERATING ENVIRONMENT Machines, OS, Network Multicast Transport Layer: TCP/IP, Wireless, Infrared, SSL	

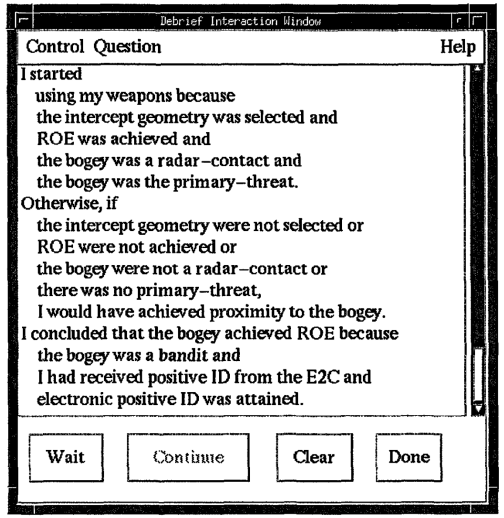
Explanation of Agent Conflicts & Harmful Interactions

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



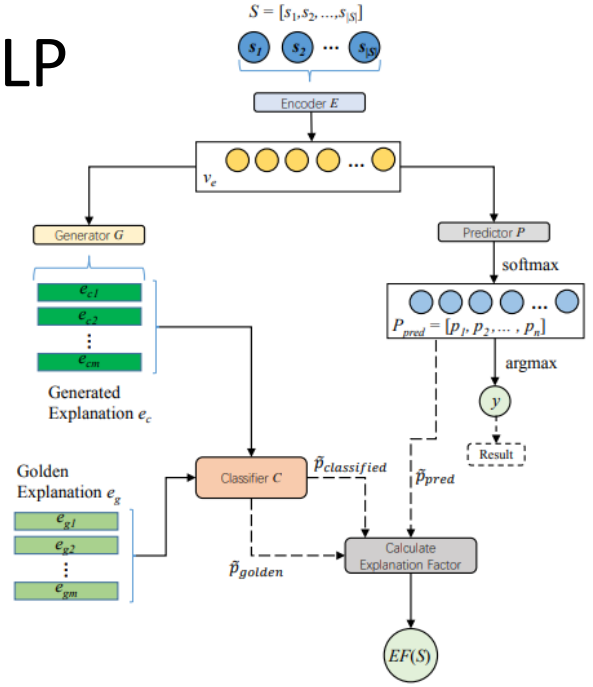
Explainable Agents

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

W. Lewis Johnson: Agents that Learn to Explain Themselves. AAAI 1994: 1257-1263

Overview of explanation in different AI fields (5)

- NLP



Fine-grained explanations are in the form of:

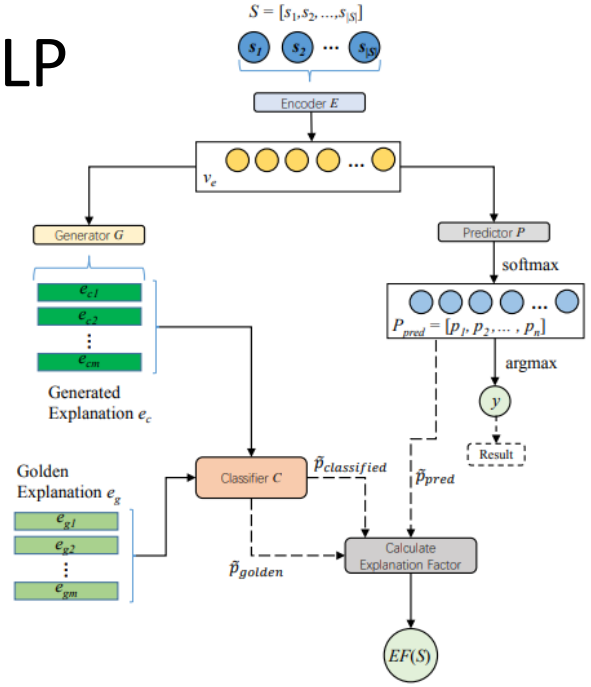
- texts in a real-world dataset;
- Numerical scores

Explainable NLP

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

Overview of explanation in different AI fields (5)

- NLP

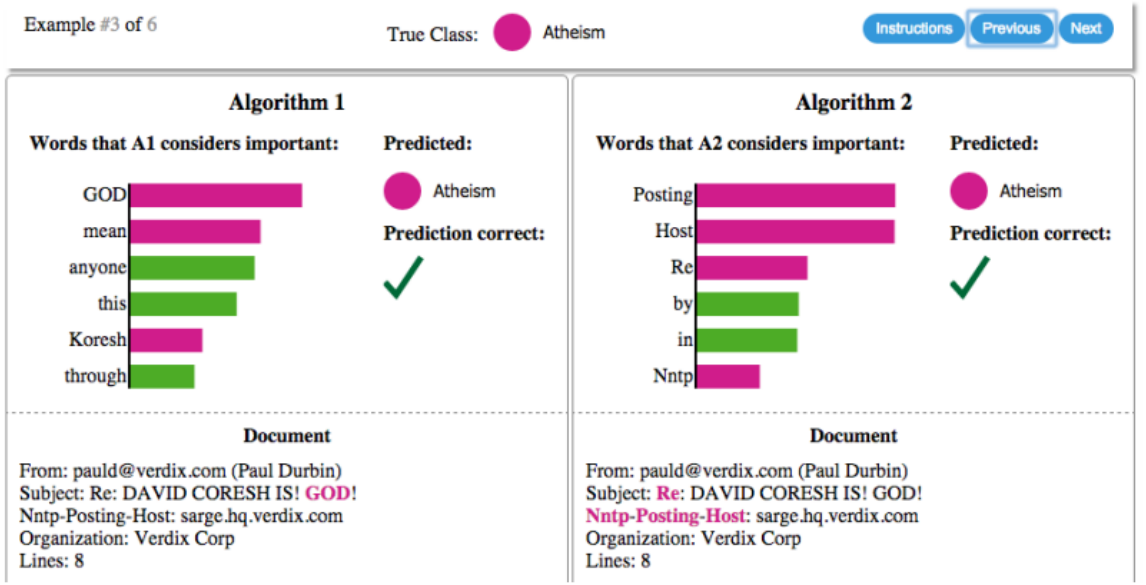


Explainable NLP

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

Fine-grained explanations are in the form of:

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

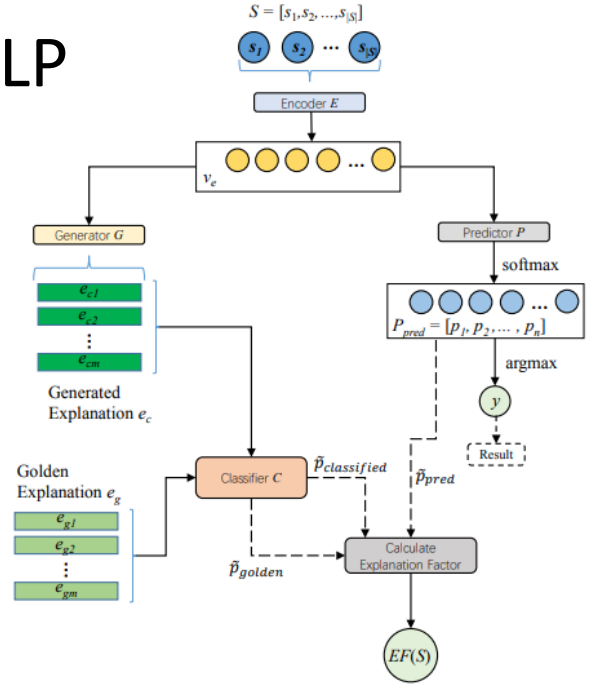


LIME for NLP

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 (5)

- NLP

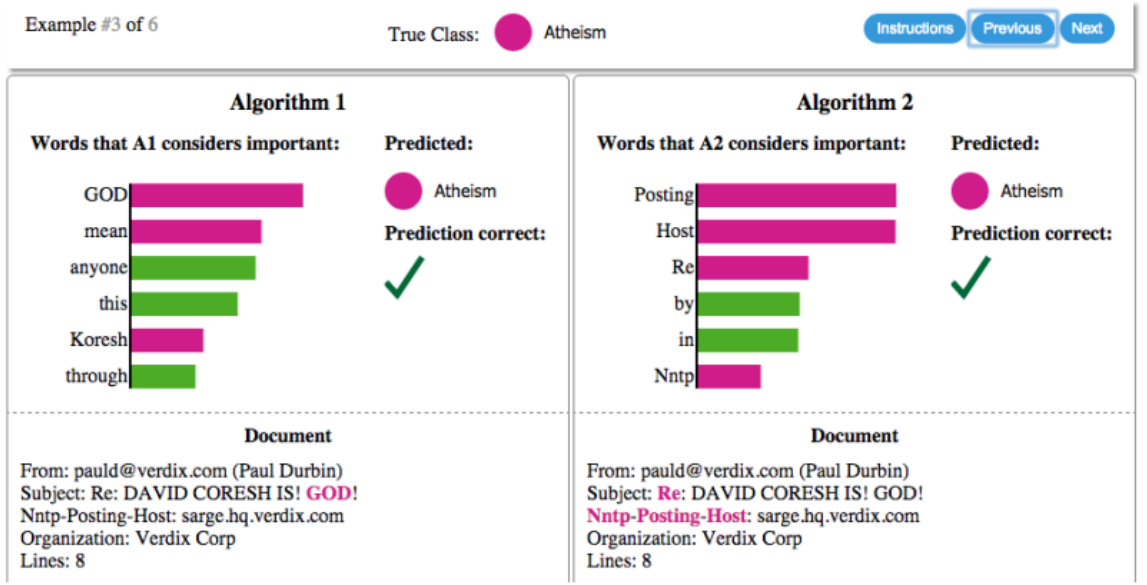


Fine-grained explanations are in the form of:

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

Explainable NLP

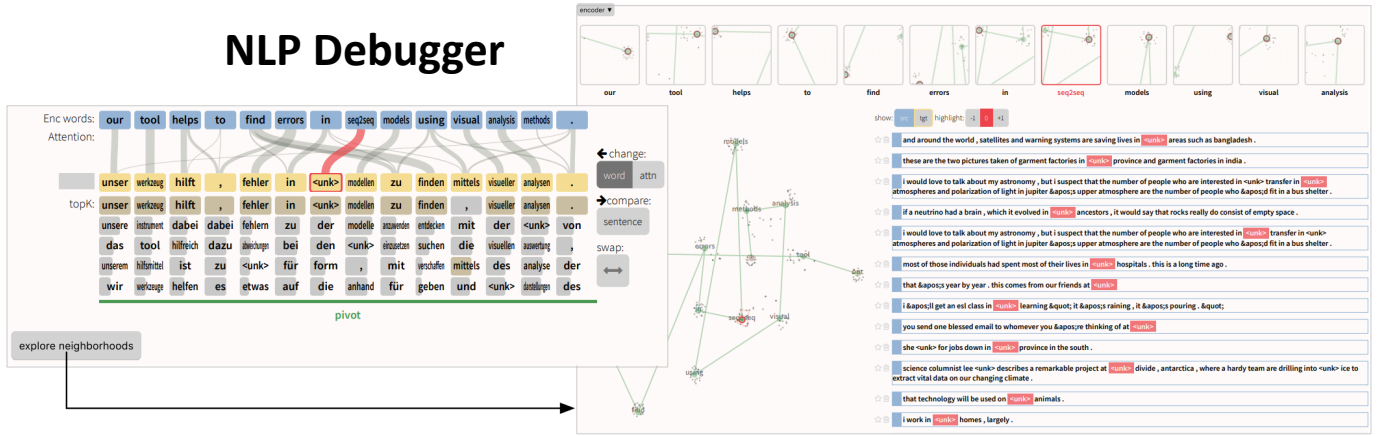
Hui Liu, Qingyu Yin, William Yang Wang: Towards Explainable NLP: A Generative Explanation Framework for Text Classification. CoRR abs/1811.00196 (2018)



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

NLP Debugger



Hendrik Strobelt, Sebastian Gehrmann, Hanspeter Pfister, Alexander M. Rush: LSTMVis: A Tool for Visual Analysis of Hidden State Dynamics in Recurrent Neural Networks. IEEE Trans. Vis. Comput. Graph. 24(1): 667-676 (2018)

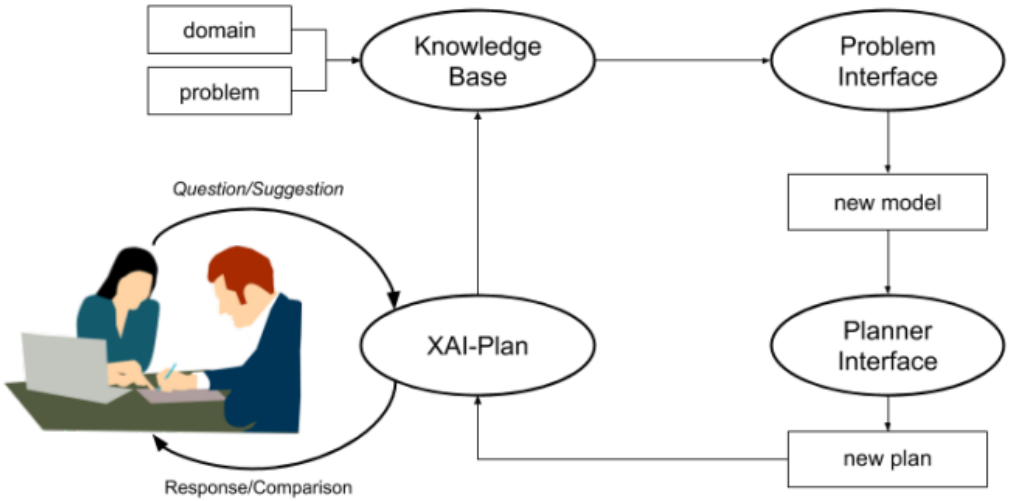
Hendrik Strobelt, Sebastian Gehrmann, Michael Behrisch, Adam Perer, Hanspeter Pfister, Alexander M. Rush: Seq2seq-Vis: A Visual Debugging Tool for Sequence-to-Sequence Models. IEEE Trans. Vis. Comput. Graph. 25(1): 353-363 (2019)

Overview of explanation in different AI fields (6)

- Planning and Scheduling

Explanation Type	R1	R2	R3	R4
Plan Patch Explanation / VAL	✗	✓	✗	✓
Model Patch Explanation	✓	✗	✓	✓
Minimally Complete Explanation	✓	✓	✗	?
Minimally Monotonic Explanation	✓	✓	✓	?
(Approximate) Minimally Complete Explanation	✗	✓	✗	✓

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



XAI Plan

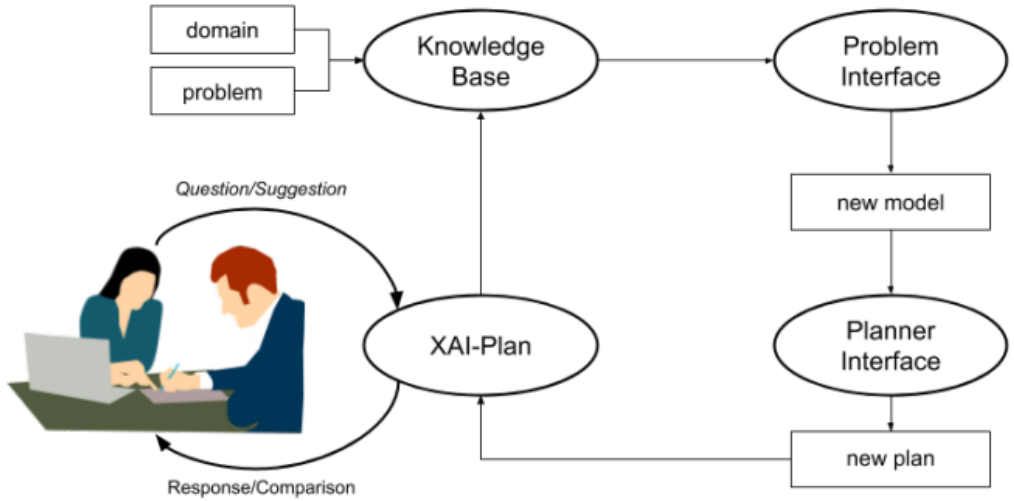
Rita Borgo, Michael Cashmore, Daniele Magazzeni: Towards Providing Explanations for AI Planner Decisions. CoRR abs/1810.06338 (2018)

Overview of explanation in different AI fields (6)

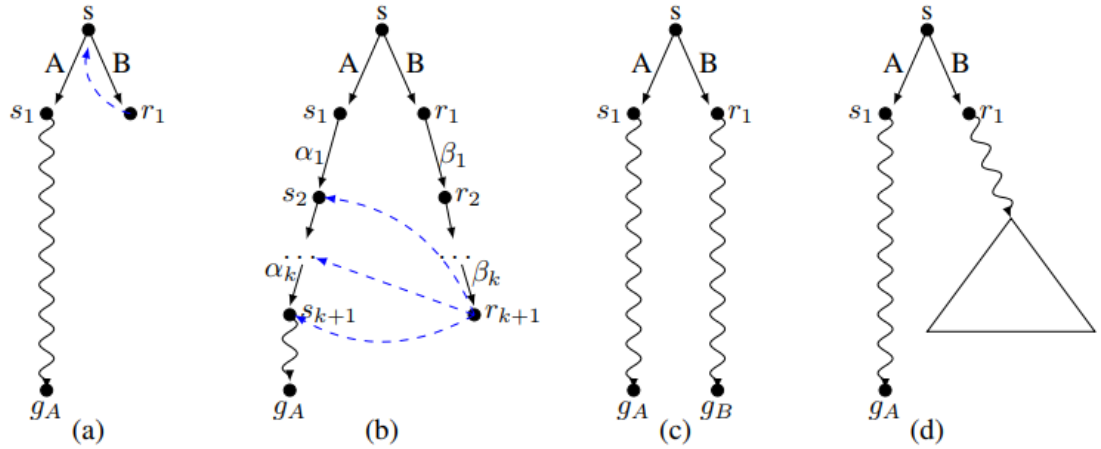
- Planning and Scheduling

Explanation Type	R1	R2	R3	R4
Plan Patch Explanation / VAL	✗	✓	✗	✓
Model Patch Explanation	✓	✗	✓	✓
Minimally Complete Explanation	✓	✓	✗	?
Minimally Monotonic Explanation	✓	✓	✓	?
(Approximate) Minimally Complete Explanation	✗	✓	✗	✓

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



XAI Plan



Human-in-the-loop Planning

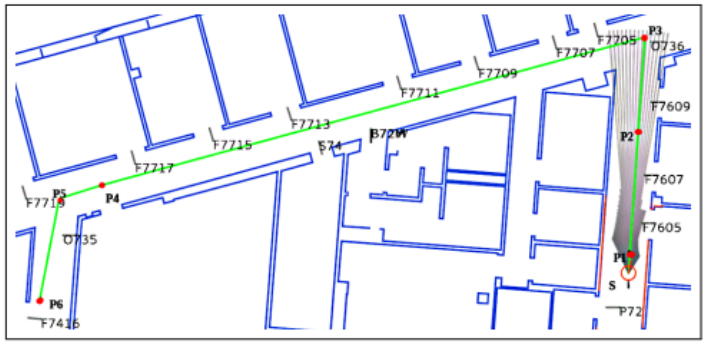
Maria Fox, Derek Long, Daniele Magazzeni: Explainable Planning. CoRR abs/1709.10256 (2017)

(Manual) Plan Comparison

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 (7)

- Robotics



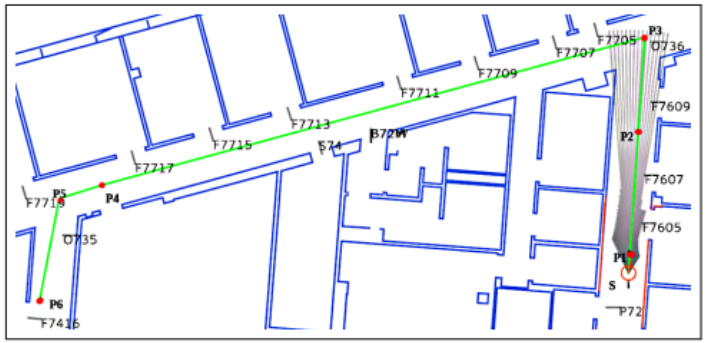
Specificity, S		Abstraction, A			
		Level 1	Level 2	Level 3	Level 4
	General Picture	Start and finish point of the complete route	Total distance and time taken for the complete route	Total distance and time taken for the complete route	Starting and ending landmark of complete route
	Summary	Start and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

Narration of Autonomous Robot Experience

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

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

Overview of explanation in different AI fields (7)



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

Narration of Autonomous Robot Experience

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

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

Robot: I have decided to turn left.

Human: Why did you do that?

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

Human: How confident are you in this decision?

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

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

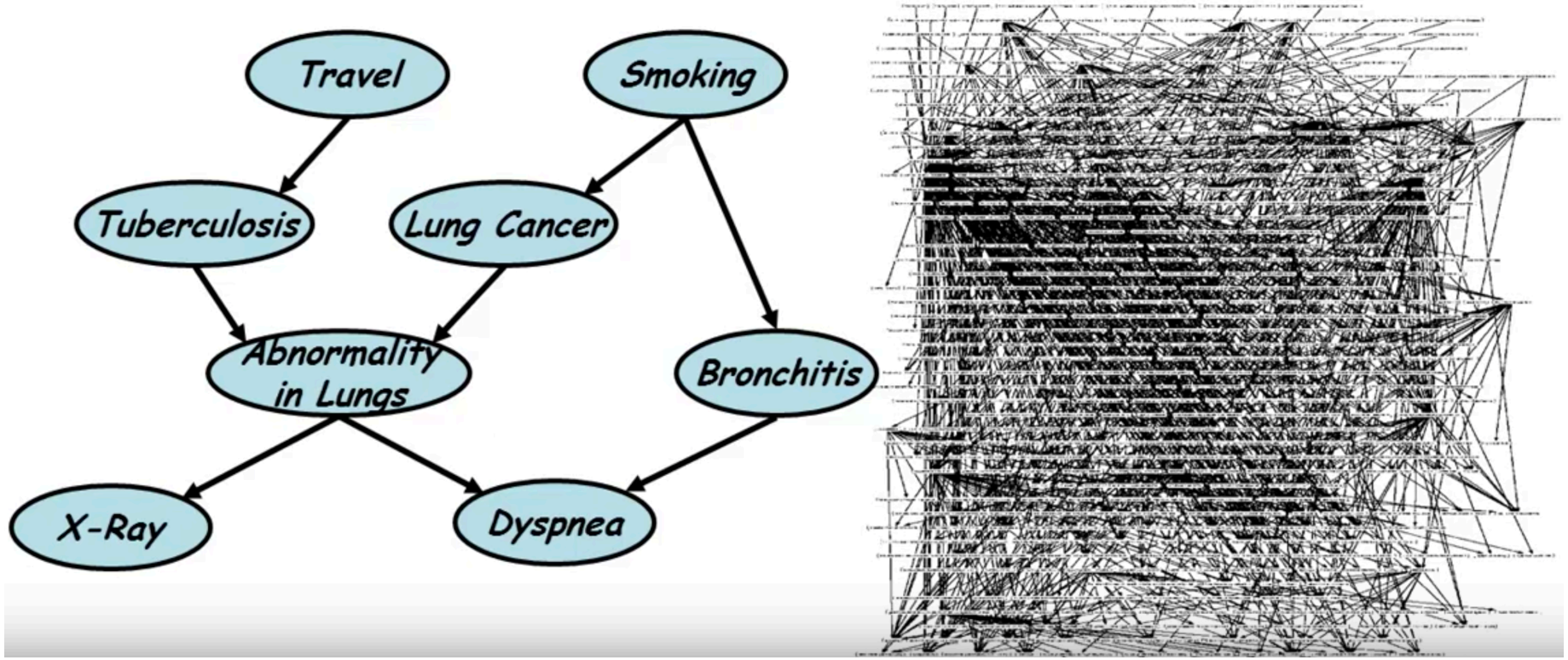
Robot: Here is the histogram of all training examples that reached this leaf. 80% of examples where this area was above 20 cm predicted the appropriate action to be "drive forward".

From Decision Tree to human-friendly information

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

Overview of explanation in different AI fields (8)

- Reasoning under Uncertainty

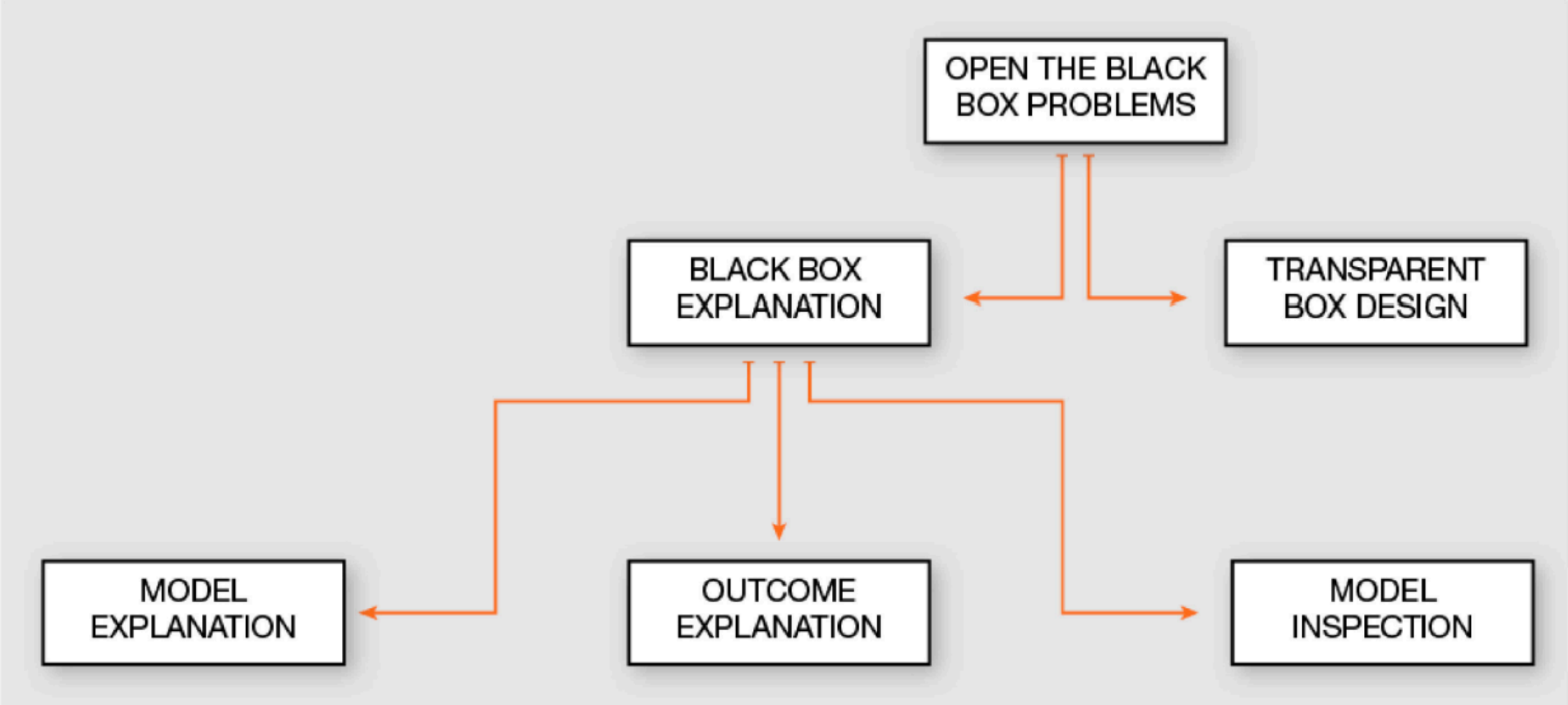


Probabilistic Graphical Models

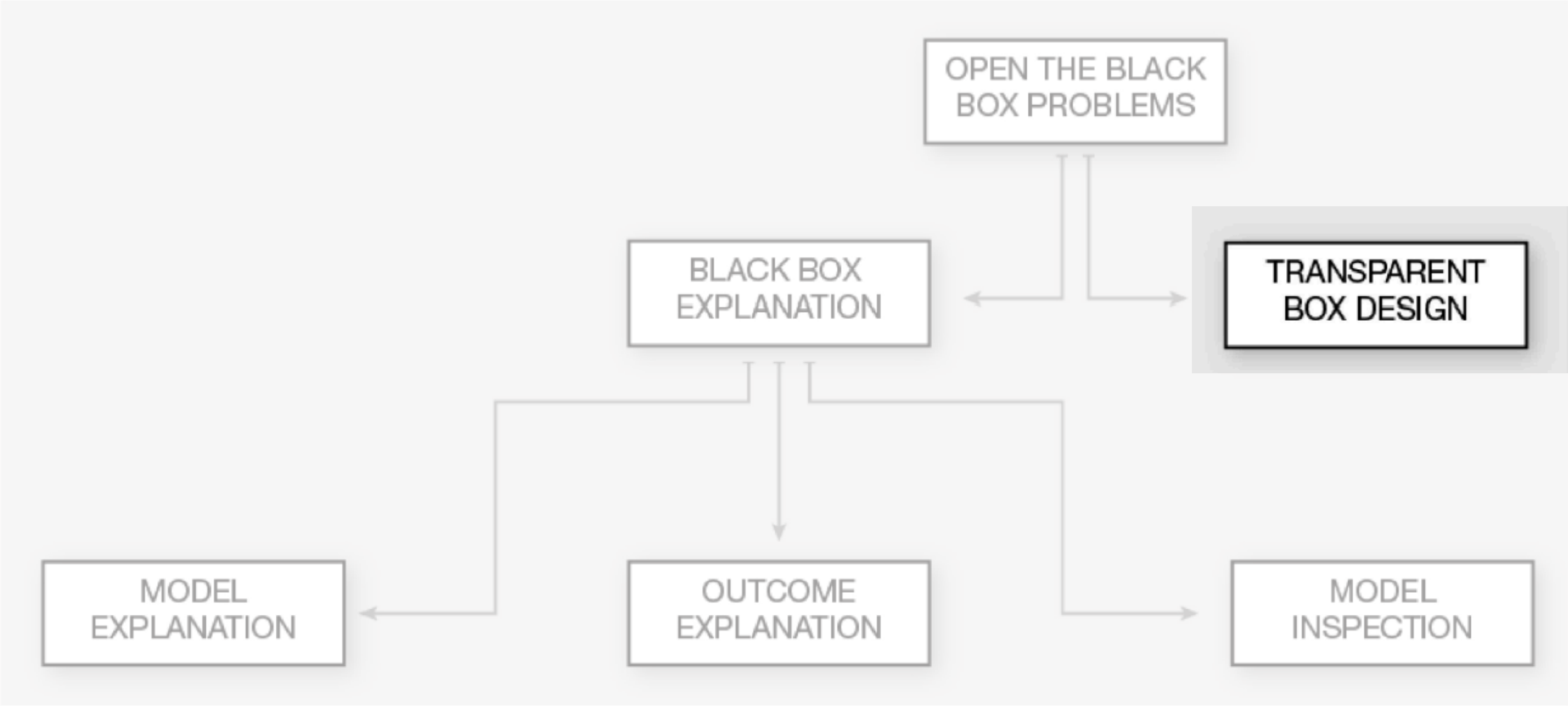
Daphne Koller, Nir Friedman: Probabilistic Graphical Models - Principles and Techniques. MIT Press 2009, ISBN 978-0-262-01319-2, pp. I-XXXV, 1-1231

XAI in Machine Learning

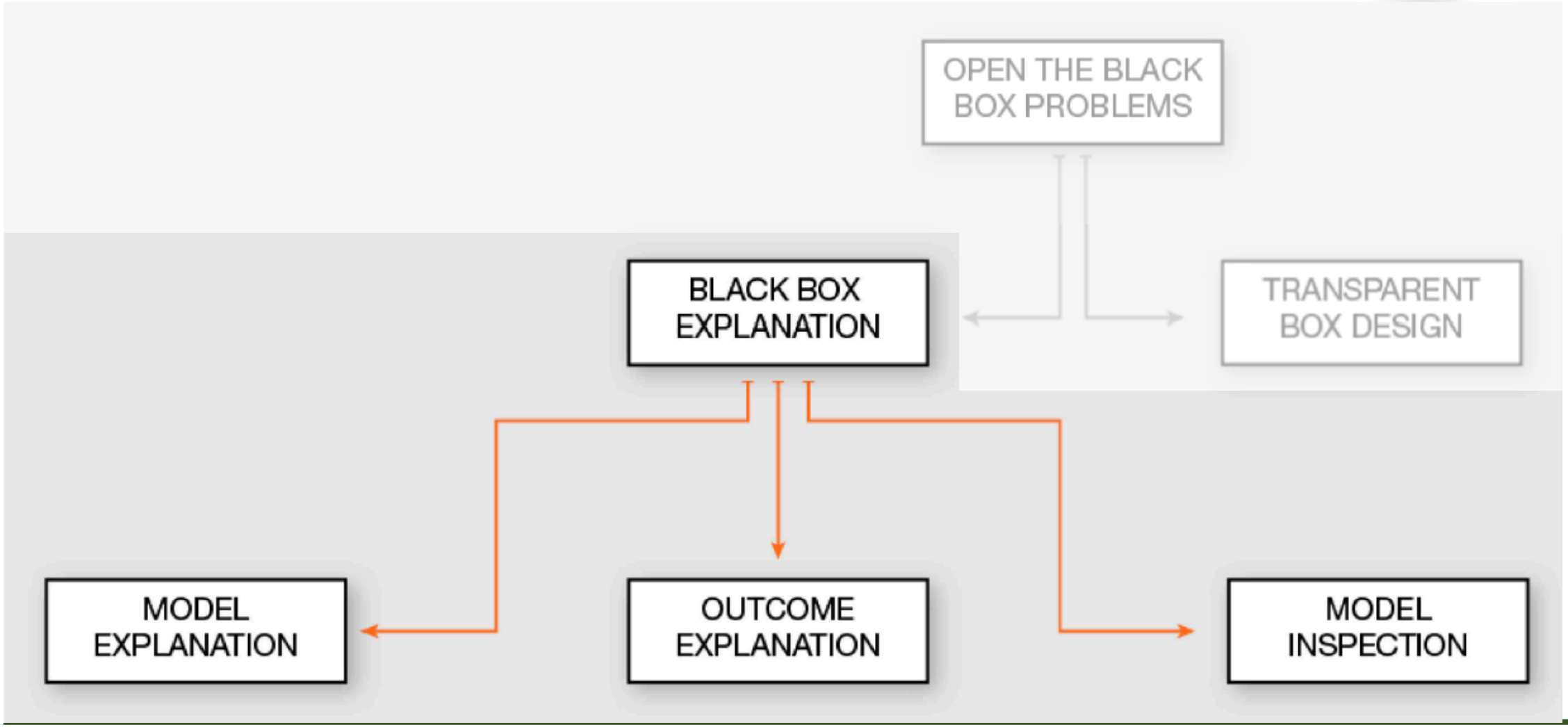
Problems Taxonomy



Explanation by Design



Black Box eXplanation

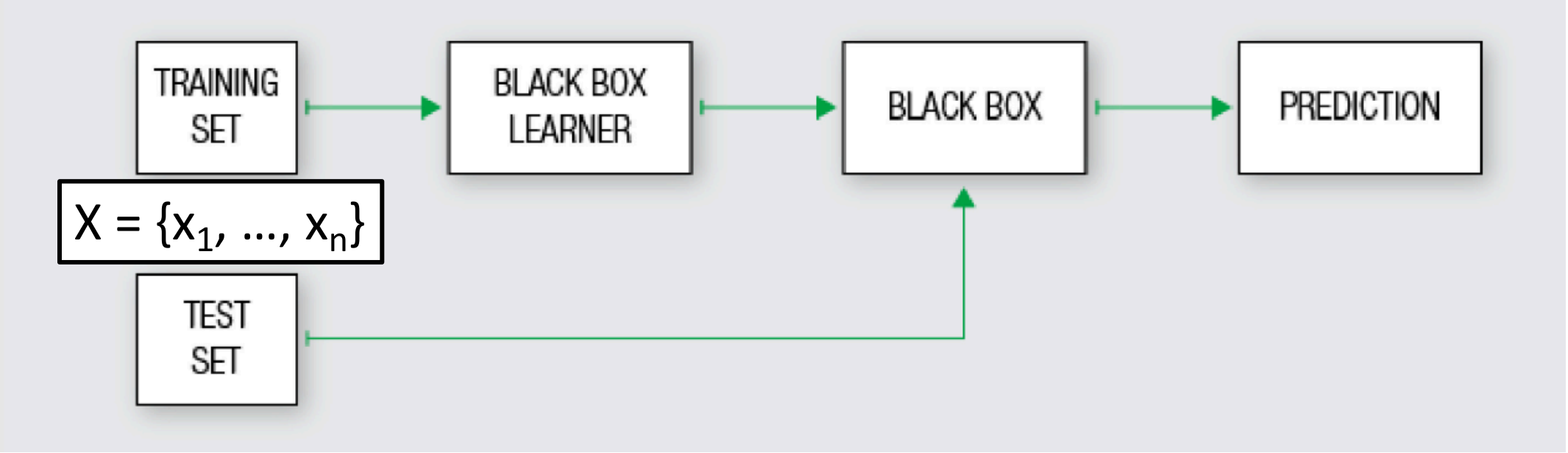


Example of XAI For a Classification Task

Guidotti, R.; Monreale, A.; Ruggieri, S.; Turini, F.; Giannotti, F.; and Pedreschi, D. 2018. A survey of methods for explaining black box models. ACM Comput. Surv. 51(5):93:1–93:42.

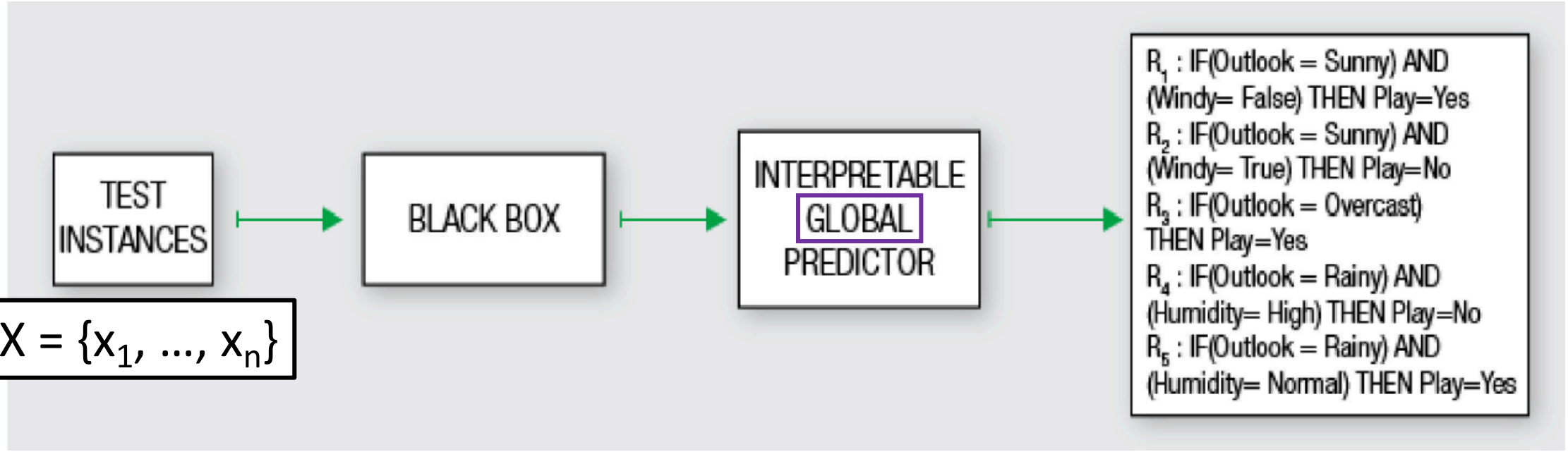
https://xaitutorial2019.github.io/

Classification Problem



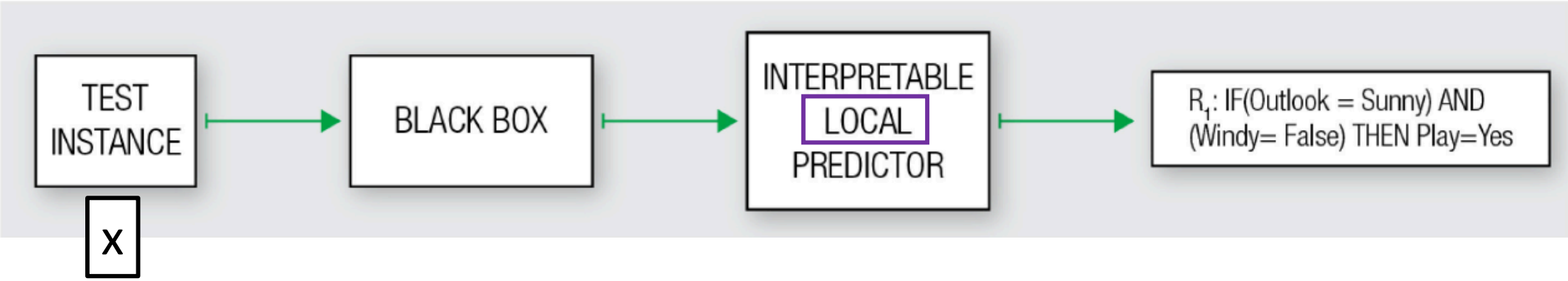
Model Explanation Problem

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



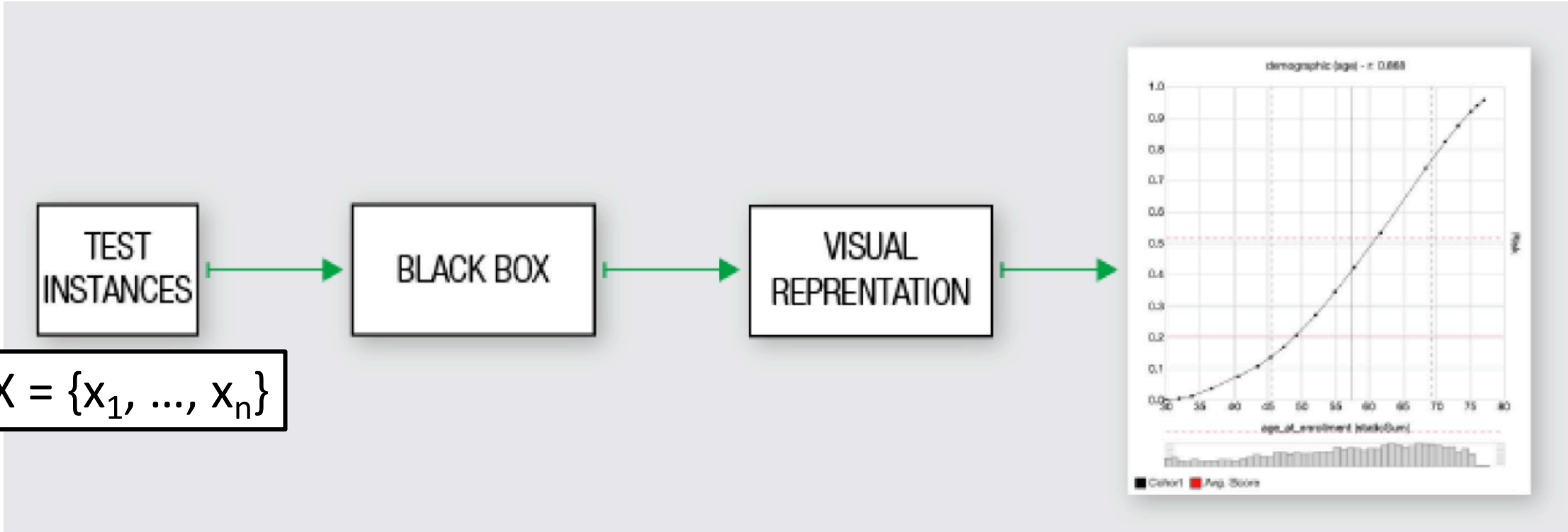
Post-hoc Explanation Problem

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

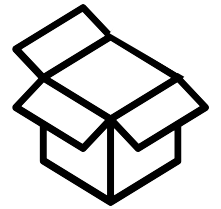


Model Inspection Problem

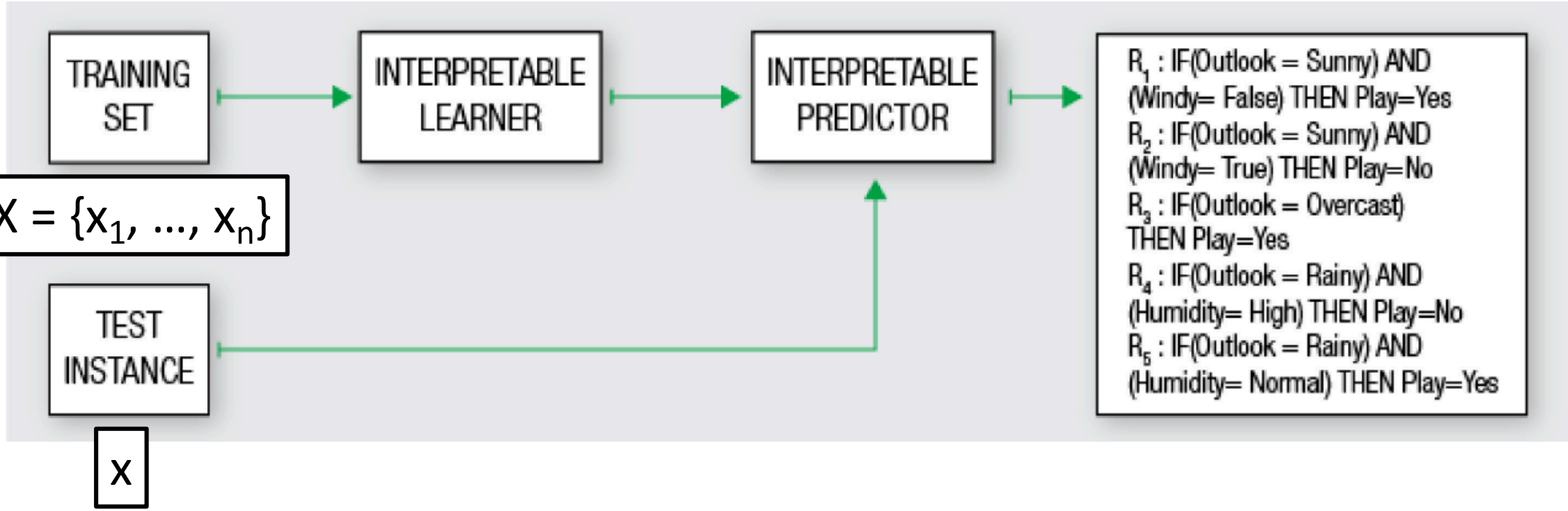
Provide a representation (visual or textual) for understanding either how the black box model works or why the black box returns certain predictions more likely than others.



Transparent Box Design Problem



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



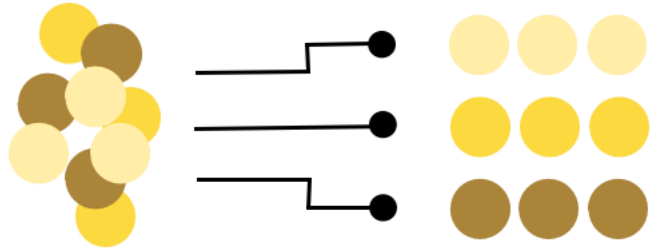
State of the Art XAI in Machine Learning

(By XAI Problem to be Solved)

Guidotti, R.; Monreale, A.; Ruggieri, S.; Turini, F.; Giannotti, F.; and Pedreschi, D. 2018. A survey of methods for explaining black box models. ACM Comput. Surv. 51(5):93:1–93:42.

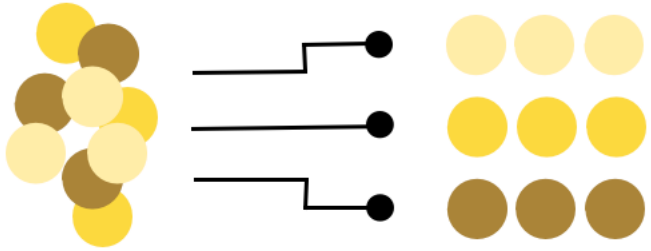
https://xaitutorial2019.github.io/

Categorization



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

Black Boxes



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

Types of Data

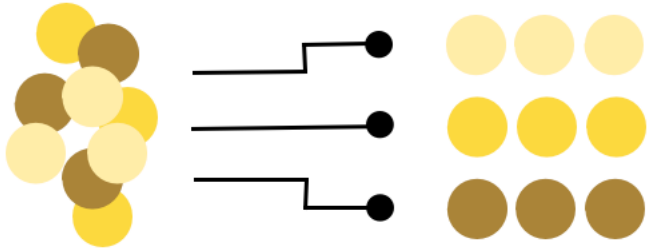


Table of baby-name data
(baby-2010.csv)

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

Field names

One row (4 fields)

2000 rows all told

Tabular

(**TAB**)

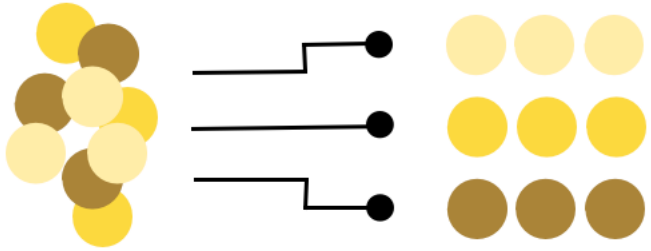
Images

(**IMG**)

Text

(**TXT**)

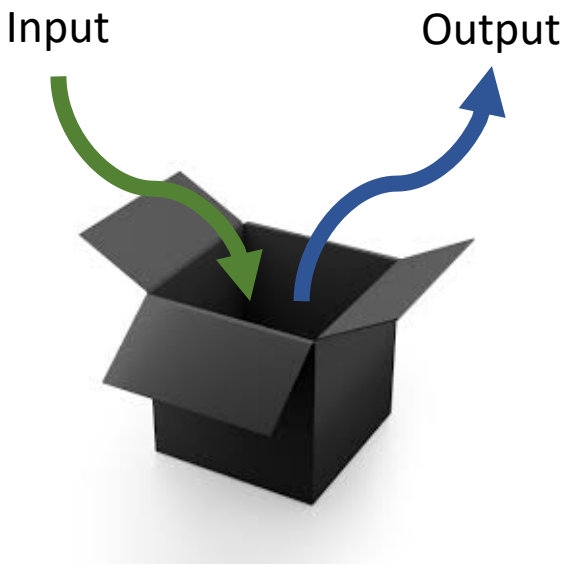
Explanators



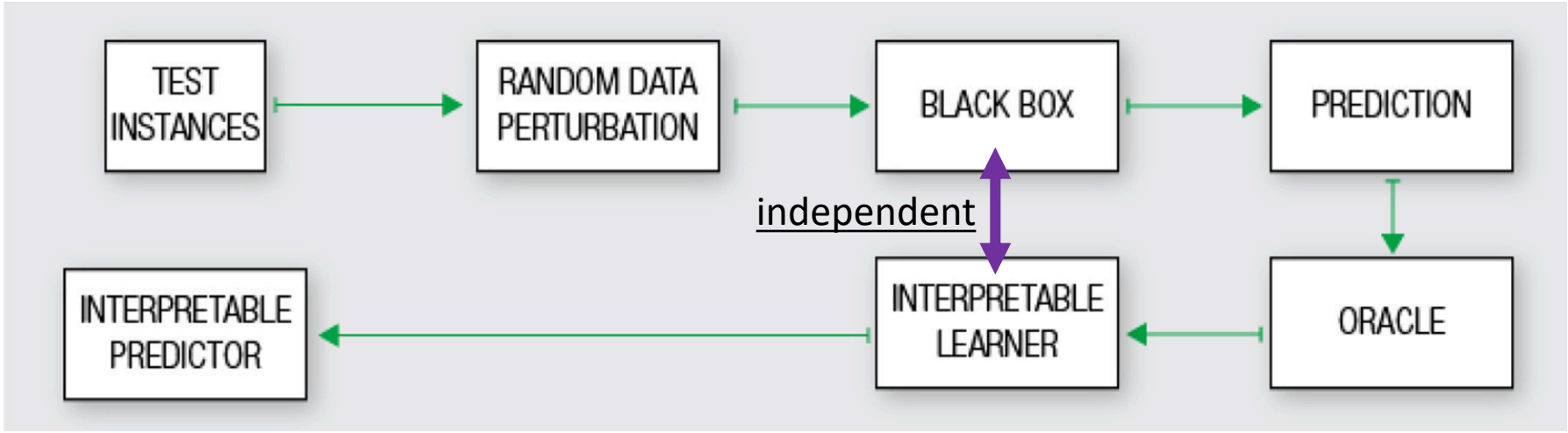
- Decision Tree (***DT***)
- Decision Rules (***DR***)
- Features Importance (***FI***)
- Saliency Mask (***SM***)
- Sensitivity Analysis (***SA***)
- Partial Dependence Plot (***PDP***)
- Prototype Selection (***PS***)
- Activation Maximization (***AM***)

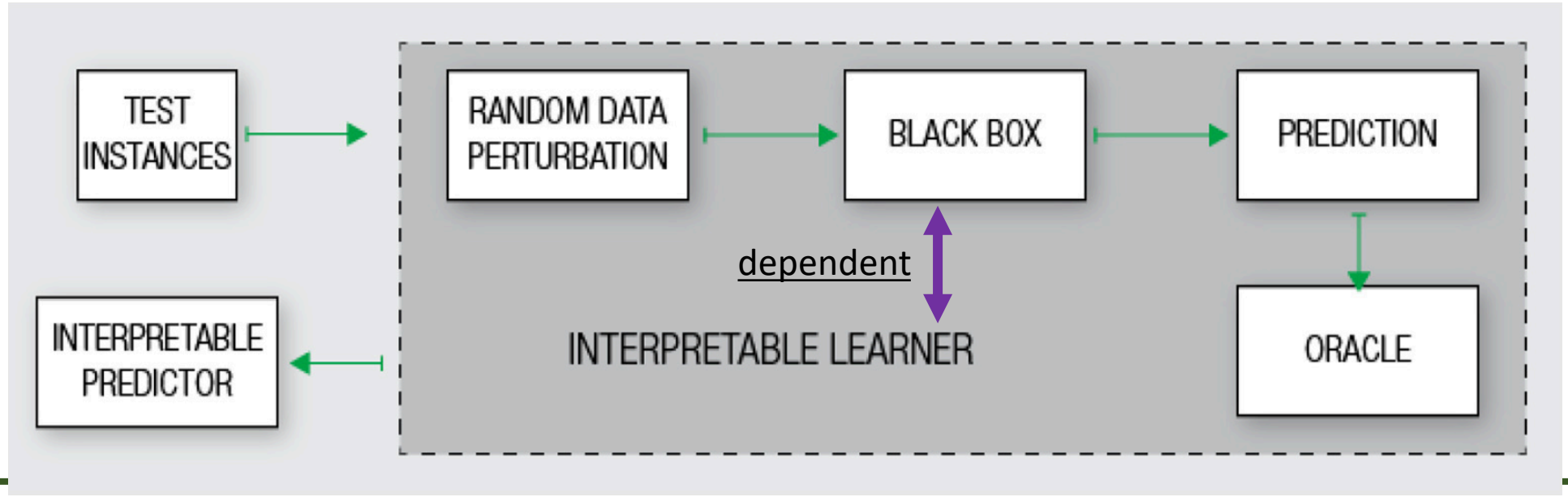
Reverse Engineering

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



Model-Agnostic vs Model-Specific





Name	Ref.	Authors	Year	Explanator	Black Box	Data Type	General	Random	Examples	Code	Dataset
Trepan	[22]	Craven et al.	1996	DT	NN	TAB	✓				✓
–	[57]	Krishnan et al.	1999	DT	NN	TAB	✓		✓		✓
DecText	[12]	Boz	2002	DT	NN	TAB	✓	✓			✓
GPDT	[46]	Johansson et al.	2009	DT	NN	TAB	✓	✓	✓		✓
Tree Metrics	[17]	Chipman et al.	1998	DT	TE	TAB					✓
CCM	[26]	Domingos et al.	1998	DT	TE	TAB	✓	✓			✓
–	[34]	Gibbons et al.	2013	DT	TE	TAB	✓	✓			
STA	[140]	Zhou et al.	2016	DT	TE	TAB		✓			
CDT	[104]	Schetinin et al.	2007	DT	TE	TAB			✓		
–	[38]	Hara et al.	2016	DT	TE	TAB		✓	✓		✓
TSP	[117]	Tan et al.	2016	DT	TE	TAB					✓
Conj Rules	[21]	Craven et al.	1995	DR	NN	TAB		✓			
G-REX	[44]	Johansson et al.	2003	DR	NN	TAB	✓	✓	✓		
REFNE	[141]	Zhou et al.	2003	DR	NN	TAB	✓	✓	✓		✓
RxREN	[6]	Augasta et al.	2012	DR	NN	TAB		✓	✓		✓

Solving The Model Explanation Problem

Guidotti, R.; Monreale, A.; Ruggieri, S.; Turini, F.; Giannotti, F.; and Pedreschi, D. 2018. A survey of methods for explaining black box models. ACM Comput. Surv. 51(5):93:1–93:42.

Global Model Explainers

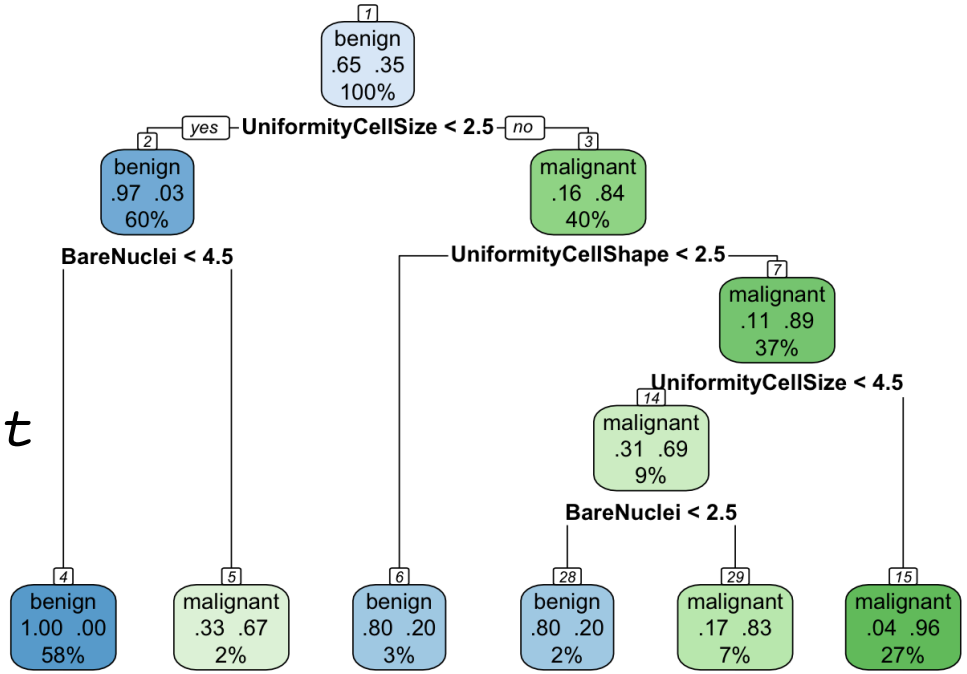
- Explanator: DT
  - Black Box: NN, TE
  - Data Type: TAB
- Explanator: DR
  - Black Box: NN, SVM, TE
  - Data Type: TAB
- Explanator: FI
  - Black Box: AGN
  - Data Type: TAB

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

Trepan – DT, NN, TAB

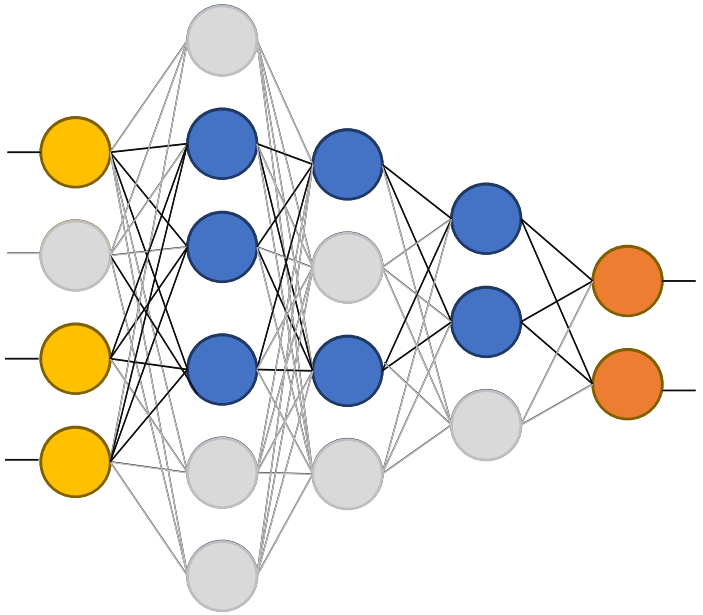
```
01    T = root_of_the_tree()
02    Q = <T, X, {}>
03    while Q not empty & size(T) < limit
04          N, X_N, C_N  = pop(Q)
05          Z_N = random(X_N, C_N)
06          y_Z = b(Z), y = b(X_N)
07          if same_class(y ∪ y_Z)
08                continue
09          S = best_split(X_N ∪ Z_N, y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

black box auditing →



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

RxREN – DR, NN, TAB



```
01    prune insignificant neurons
02    for each significant neuron
03      for each outcome
04        compute mandatory data ranges
05    for each outcome
06      build rules using data ranges of each neuron
07    prune insignificant rules
08    update data ranges in rule conditions analyzing error
```

black box auditing →

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

else

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

else

class $= C_2$

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

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

Solving The Outcome Explanation Problem

Guidotti, R.; Monreale, A.; Ruggieri, S.; Turini, F.; Giannotti, F.; and Pedreschi, D. 2018. A survey of methods for explaining black box models. ACM Comput. Surv. 51(5):93:1–93:42.

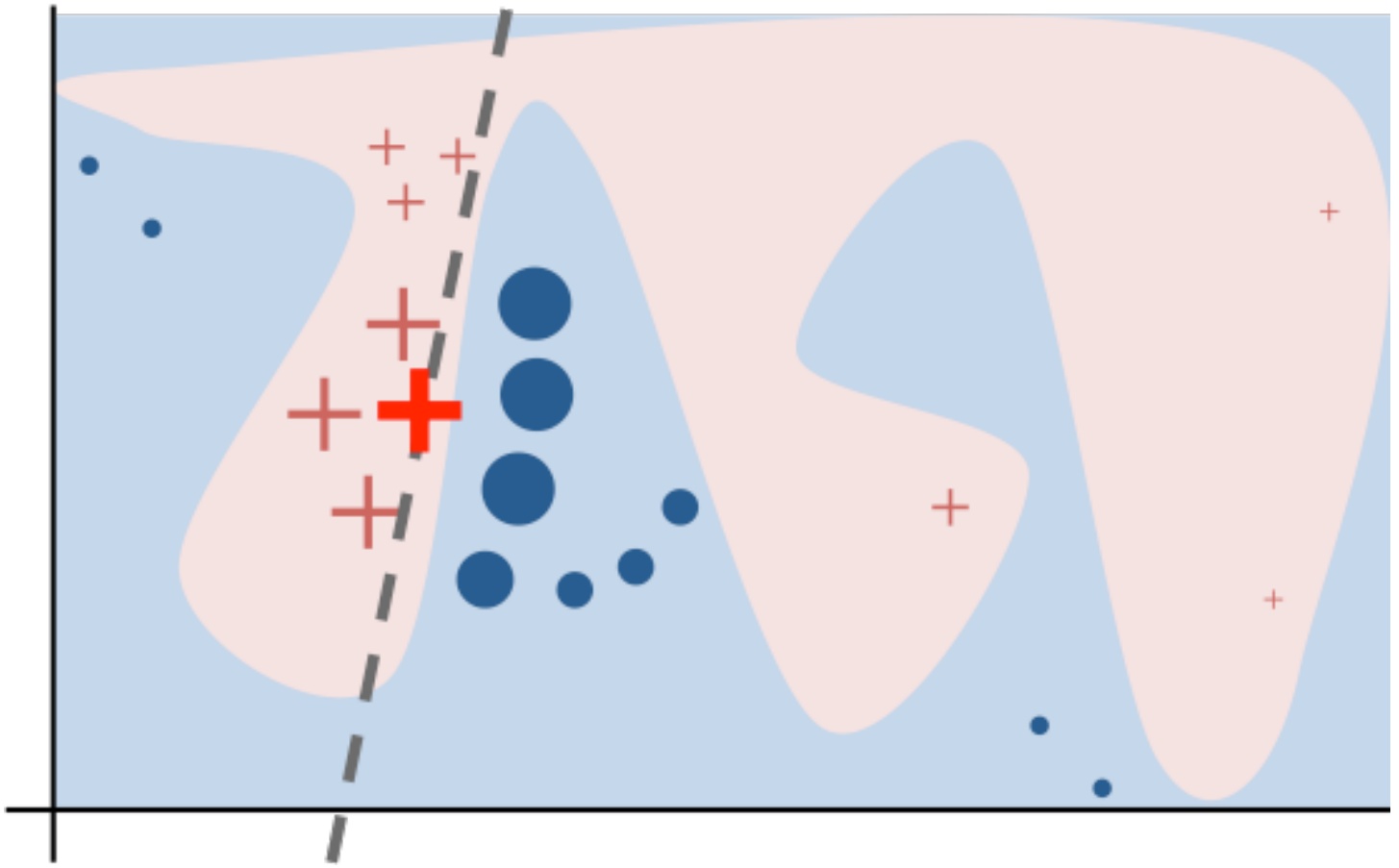
Local Model Explainers

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

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

Local Explanation

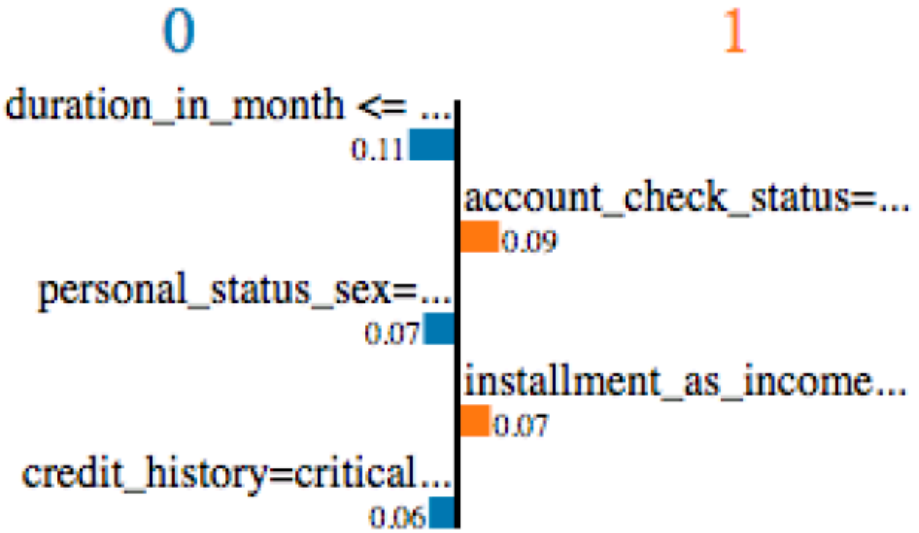
- The overall decision boundary is complex
- In the neighborhood of a single decision, the boundary is simple
- A single decision can be explained by auditing the black box around the given instance and learning a ***local*** decision.

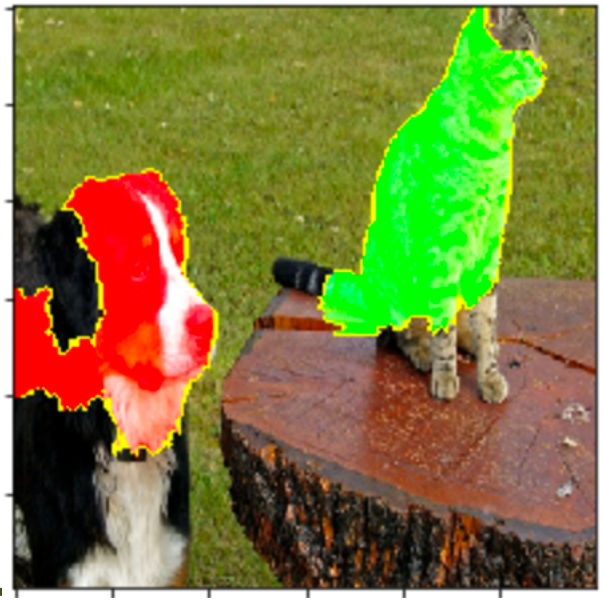


LIME – FI, AGN, ANY

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

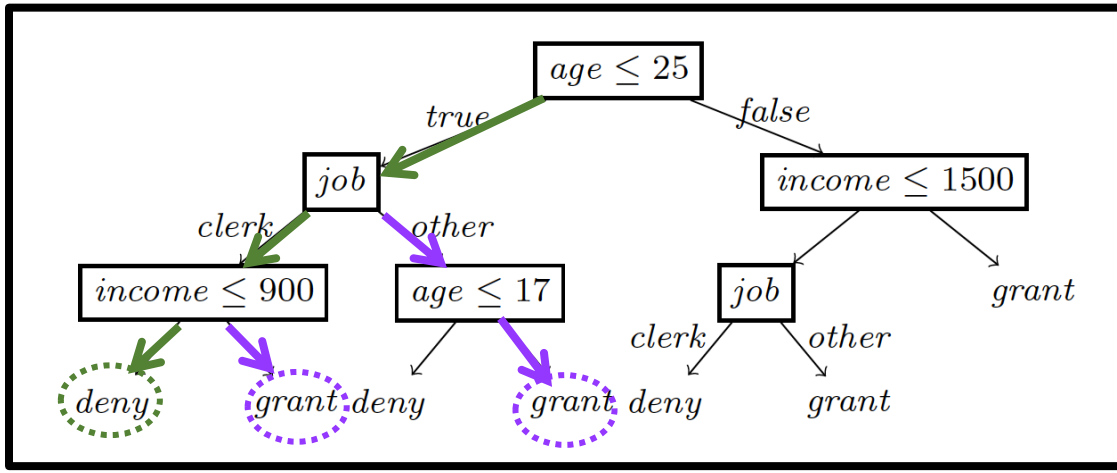
black box auditing





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

LORE – DR, AGN, TAB



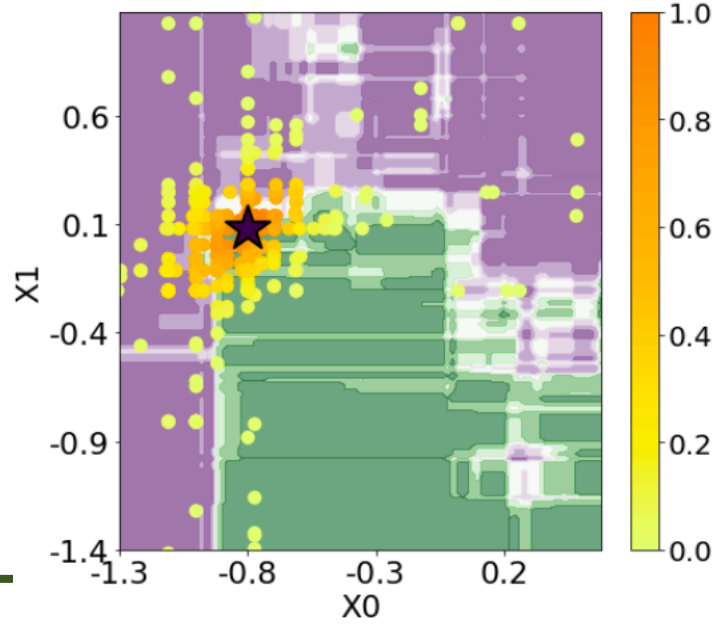

```
01    x instance to explain
02    Z_= = geneticNeighborhood(x, fitness_=, N/2)
03    Z_≠ = geneticNeighborhood(x, fitness_≠, N/2)
04    Z = Z_= ∪ Z_≠
05    c = buildTree(Z, b(Z))
06    r = (p -> y) = extractRule(c, x)
07    ϕ = extractCounterfactual(c, r, x)
08    return e = <r, ϕ>
```

black box auditing

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

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

Pedreschi, Franco Turini,
f black box decision



Meaningful Perturbations – SM, DNN, IMG

01 x instance to explain

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

black box auditing

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

constant value, noise, blurred image

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

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

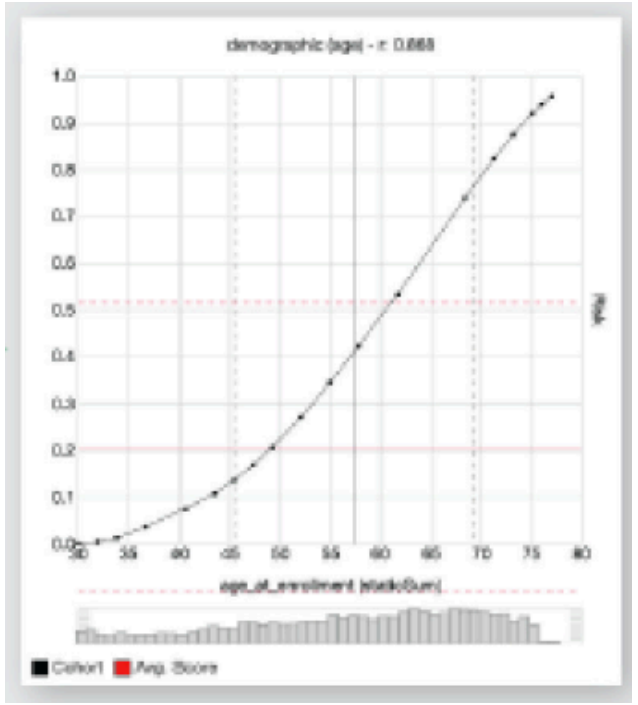
Name	Ref.	Authors	Year	Explanator	Black Box	Data Type	General	Random	Examples	Code	Dataset
NID	[83]	Olden et al.	2002	SA	NN	TAB			✓		
GDP	[8]	Baehrens	2010	SA	AGN	TAB	✓		✓		✓
QII	[24]	Datta et al	2016	SA	AGN	TAB	✓		✓		✓
IG	[115]	Sundararajan	2017	SA	DNN	ANY			✓		✓
VEC	[18]	Cortez et al.	2011	SA	AGN	TAB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

Solving The Model Inspection Problem

Guidotti, R.; Monreale, A.; Ruggieri, S.; Turini, F.; Giannotti, F.; and Pedreschi, D. 2018. A survey of methods for explaining black box models. ACM Comput. Surv. 51(5):93:1–93:42.

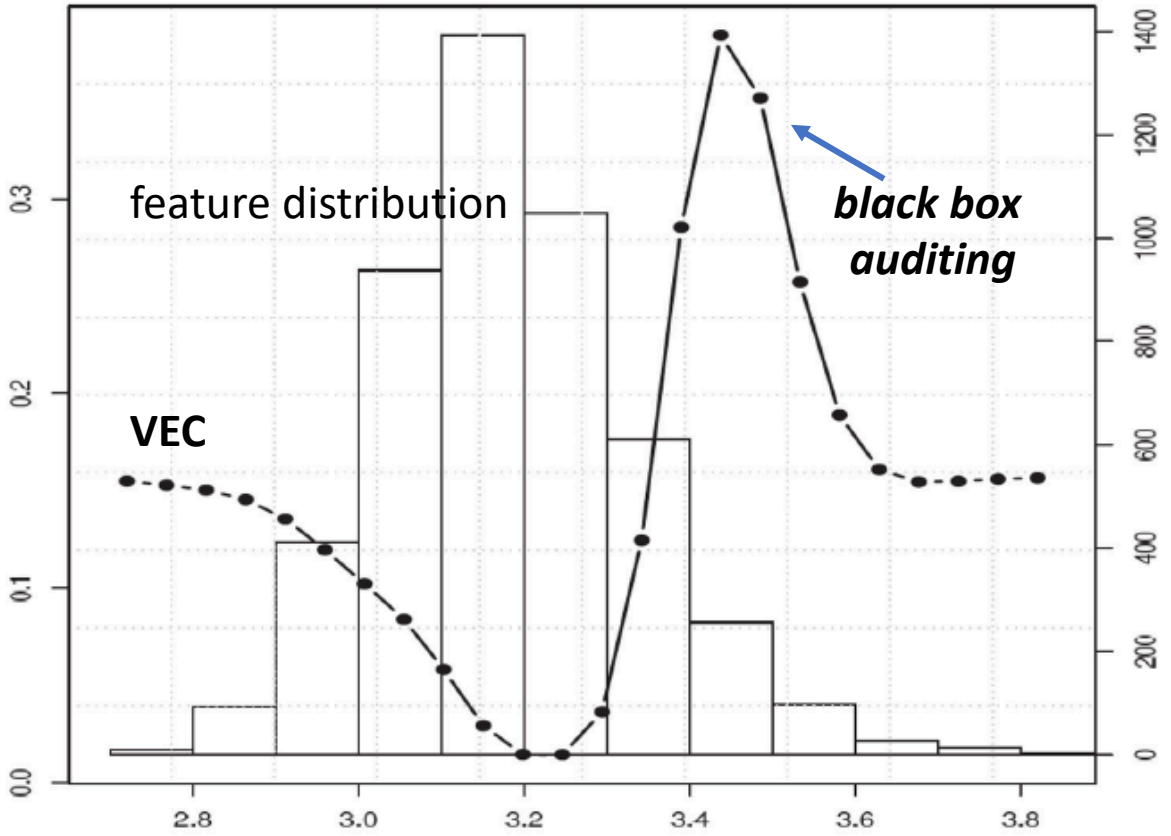
Inspection Model Explainers

- Explanator: SA
  - Black Box: NN, DNN, AGN
  - Data Type: TAB
- Explanator: PDP
  - Black Box: AGN
  - Data Type: TAB
- Explanator: AM
  - Black Box: DNN
  - Data Type: IMG, TXT



VEC – SA, AGN, TAB

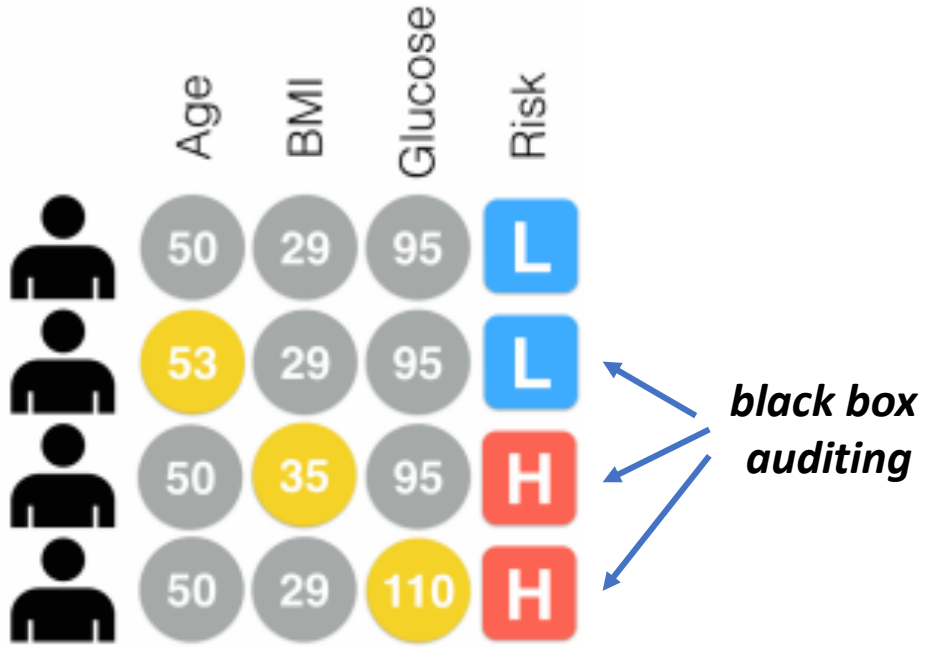
- Sensitivity measures are variables calculated as the range, gradient, variance of the prediction.
- The visualizations realized are barplots for the features importance, and ***Variable Effect Characteristic*** curve (VEC) plotting the input values versus the (average) outcome responses.

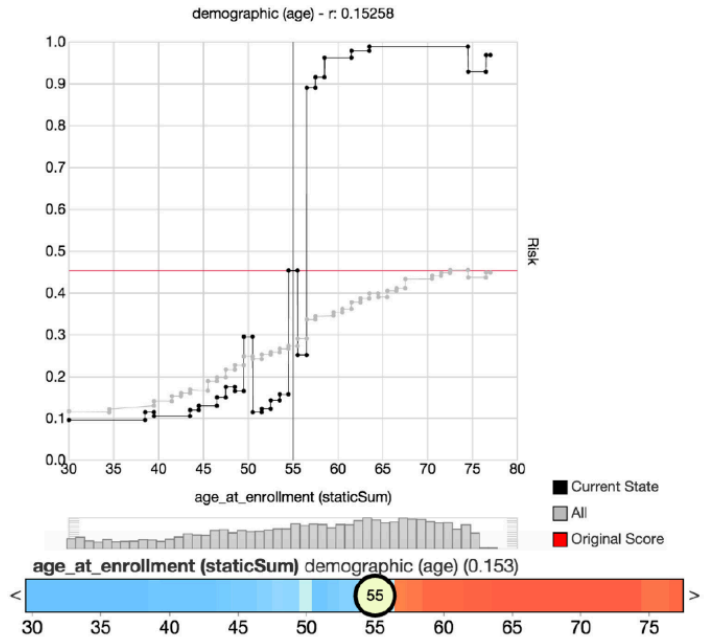


- Paulo Cortez and Mark J. Embrechts. 2011. ***Opening black box data mining models using sensitivity analysis***. CIDM.

Prospector – PDP, AGN, TAB

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





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

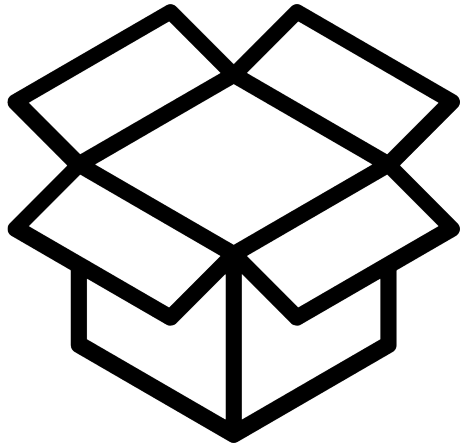
Name	Ref.	Authors	Year	Explanator	Black Box	Data Type	General	Random	Examples	Code	Dataset
CPAR	[135]	Yin et al.	2003	DR	–	TAB					✓
FRL	[127]	Wang et al.	2015	DR	–	TAB			✓	✓	✓
BRL	[66]	Letham et al.	2015	DR	–	TAB			✓		
TLBR	[114]	Su et al.	2015	DR	–	TAB			✓		✓
IDS	[61]	Lakkaraju et al.	2016	DR	–	TAB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

Solving The Transparent Design Problem

Guidotti, R.; Monreale, A.; Ruggieri, S.; Turini, F.; Giannotti, F.; and Pedreschi, D. 2018. A survey of methods for explaining black box models. ACM Comput. Surv. 51(5):93:1–93:42.

Transparent Model Explainers

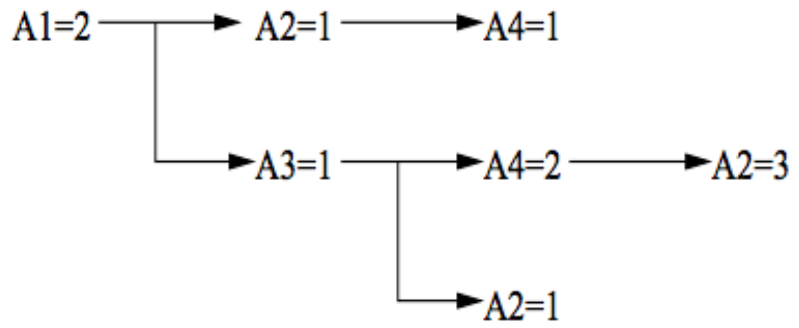
- Explanators:
  - DR
  - DT
  - PS
- Data Type:
  - TAB



CPAR – DR, TAB

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

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



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

CORELS – DR, TAB

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

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

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

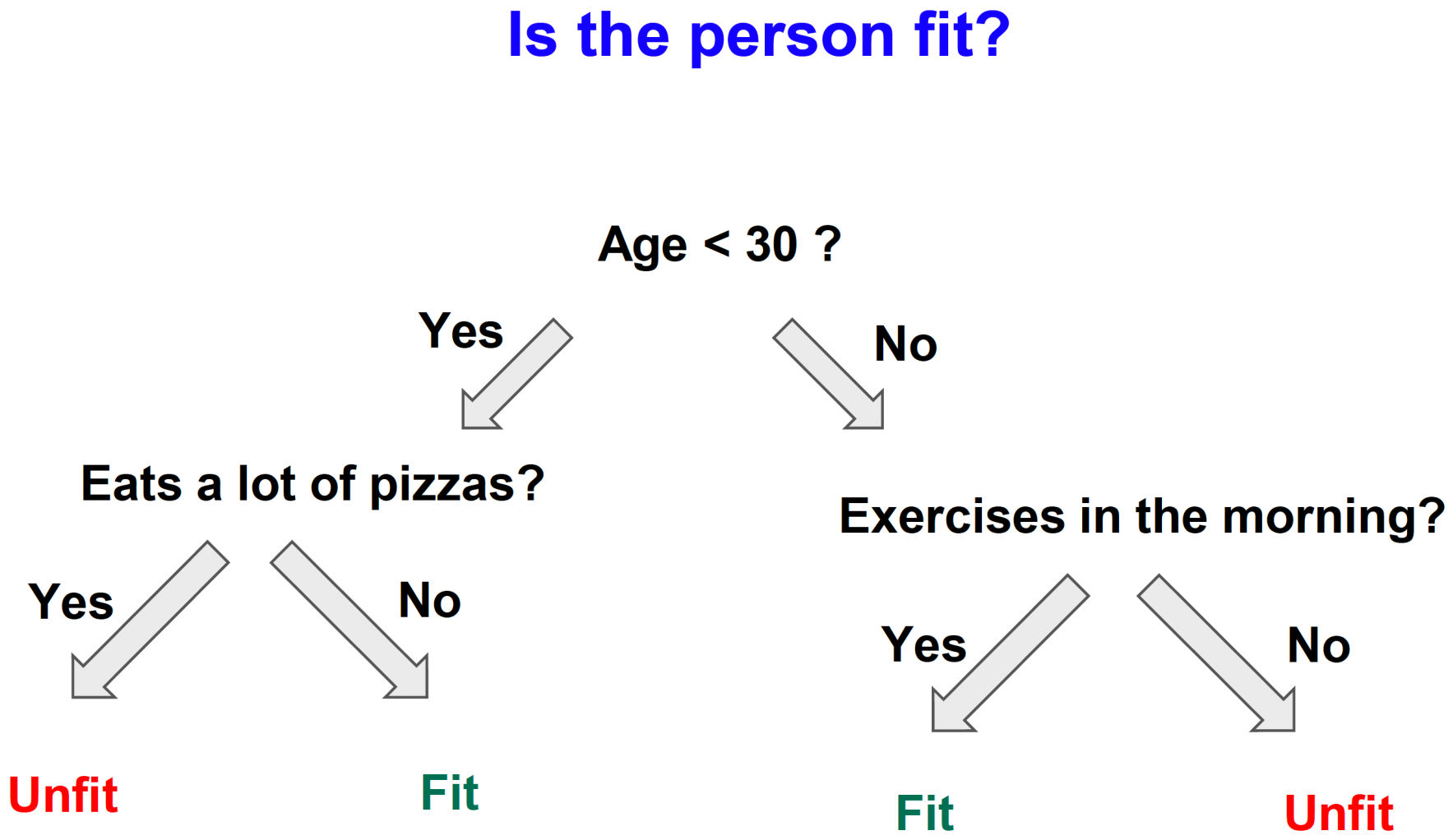
State of the Art XAI in Machine Learning

(By Machine Learning Type)

Overview of explanation in Machine Learning (1)

- All except Artificial Neural Network

Interpretable Models:
- Decision Trees



Overview of explanation in Machine Learning (1)

- All except Artificial Neural Network

Interpretable Models:

- Decision Trees, Lists

If Past-Respiratory-Illness =Yes **and** Smoker =Yes **and** Age $\geq$ 50, **then** Lung Cancer

Else if Allergies =Yes **and** Past-Respiratory-Illness =Yes, **then** Asthma

Else if Family-Risk-Respiratory =Yes, **then** Asthma

Else if Family-Risk-Depression =Yes, **then** Depression

Else if Gender =Female **and** Short-Breath-Symptoms =Yes, **then** Asthma

Else if BMI $\geq$ 0.2 **and** Age$\geq$ 60, **then** Diabetes

Else if Frequent-Headaches =Yes **and** Dizziness =Yes, **then** Depression

Else if Frequency-Doctor-Visits $\geq$ 0.3, **then** Diabetes

Else if Disposition-Tiredness =Yes, **then** Depression

Else if Chest-Pain =Yes **and** Nausea **and** Yes, **then** Diabetes

Else Diabetes

Interpretable Decision Sets: A Joint Framework for Description and Prediction, Lakkaraju, Bach, Leskovec

Overview of explanation in Machine Learning (1)

- All except Artificial Neural Network

Interpretable Models:

- Decision Trees, Lists and Sets,

If Allergies =Yes **and** Smoker =Yes **and** Irregular-Heartbeat =Yes, **then** Asthma

If Allergies =Yes **and** Past-Respiratory-Illness =Yes **and** Avg-Body-Temperature $\geq$ 0.1, **then** Asthma

If Smoker =Yes **and** BMI $\geq$ 0.2 **and** Age $\geq$ 60, **then** Diabetes

If Family-Risk-Diabetes =Yes **and** BMI $\geq$ 0.4 =Frequency-Infections $\geq$ 0.2, **then** Diabetes

If Frequency-Doctor-Visits $\geq$ 0.4 **and** Childhood-Obesity =Yes **and** Past-Respiratory-Illness =Yes, **then** Diabetes

If Family-Risk-Depression =Yes **and** Past-Depression =Yes **and** Gender =Female, **then** Depression

If BMI $\geq$ 0.3 **and** Insurance-Coverage =None **and** Avg-Blood-Pressure $\geq$ 0.2, **then** Depression

If Past-Respiratory-Illness =Yes **and** Age $\geq$ 50 **and** Smoker =Yes, **then** Lung Cancer

If Family-Risk-LungCancer =Yes **and** Allergies =Yes **and** Avg-Blood-Pressure $\geq$ 0.3, **then** Lung Cancer

If Disposition-Tiredness =Yes **and** Past-Anemia =Yes **and** BMI $\geq$ 0.3 **and** Rapid-Weight-Loss =Yes, **then** Leukemia

If Family-Risk-Leukemia =Yes **and** Past-Blood-Clotting =Yes **and** Frequency-Doctor-Visits $\geq$ 0.3, **then** Leukemia

If Disposition-Tiredness =Yes **and** Irregular-Heartbeat =Yes **and** Short-Breath-Symptoms =Yes **and** Abdomen-Pains =Yes, **then** Myelofibrosis

Interpretable Decision Sets: A Joint Framework for Description and Prediction, Lakkaraju, Bach, Leskovec

Overview of explanation in Machine Learning (1)

- All except Artificial Neural Network

Interpretable Models:

- Decision Trees, Lists and Sets,
- GAMs,
- GLMs,

Model	Form	Intelligibility	Accuracy
Linear Model	$y = \beta_0 + \beta_1 x_1 + ... + \beta_n x_n$	+++	+
Generalized Linear Model	$g(y) = \beta_0 + \beta_1 x_1 + ... + \beta_n x_n$	+++	+
Additive Model	$y = f_1(x_1) + ... + f_n(x_n)$	++	++
Generalized Additive Model	$g(y) = f_1(x_1) + ... + f_n(x_n)$	++	++
Full Complexity Model	$y = f(x_1, ..., x_n)$	+	+++

Intelligible Models for Classification and Regression. Lou, Caruana and Gehrke KDD 2012

Accurate Intelligible Models with Pairwise Interactions. Lou, Caruana, Gehrke and Hooker. KDD 2013

Overview of explanation in Machine Learning (1)

- All except Artificial Neural Network

Interpretable Models:

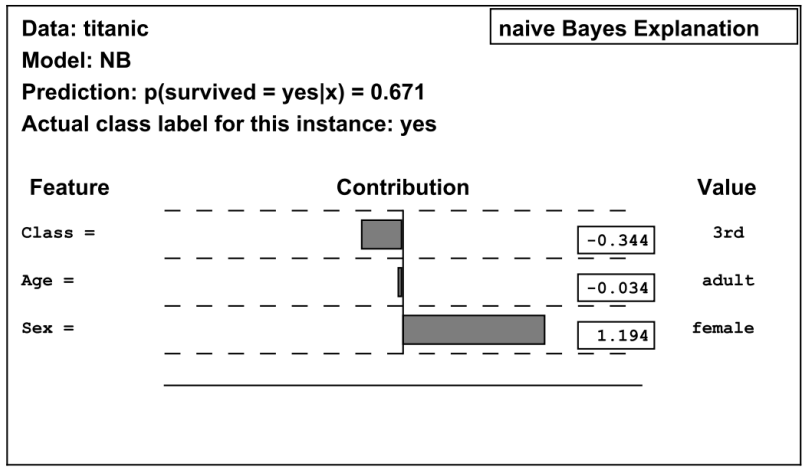
- Decision Trees, Lists and Sets,
- GAMs,
- GLMs,
- Linear regression,
- Logistic regression,
- KNNs

Overview of explanation in Machine Learning (1)

- All except Artificial Neural Network

Interpretable Models:

- Decision Trees, Lists and Sets,
- GAMs,
- GLMs,
- Linear regression,
- Logistic regression,
- KNNs



Naive Bayes model

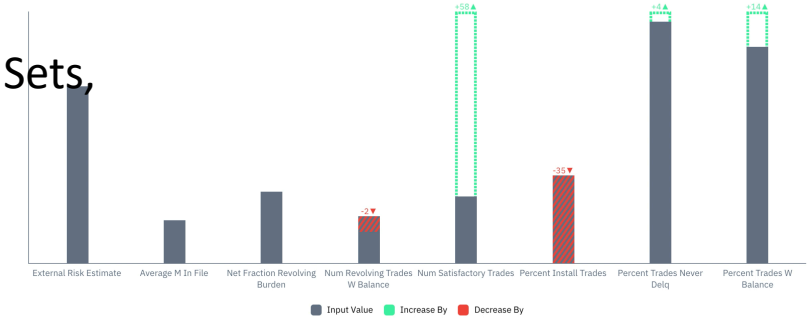
Igor Kononenko. Machine learning for medical diagnosis: history, state of the art and perspective. Artificial Intelligence in Medicine, 23:89–109, 2001.

Overview of explanation in Machine Learning (1)

- All except Artificial Neural Network

Interpretable Models:

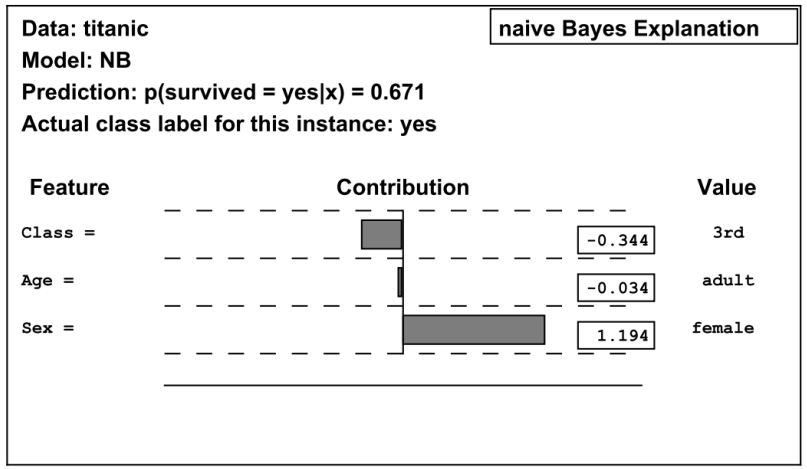
- Decision Trees, Lists and Sets,
- GAMs,
- GLMs,
- Linear regression,
- Logistic regression,
- KNNs



Counterfactual What-if

Brent D. Mittelstadt, Chris Russell, Sandra Wachter: Explaining Explanations in AI. FAT 2019: 279-288

Rory Mc Grath, Luca Costabello, Chan Le Van, Paul Sweeney, Farbod Kamiab, Zhao Shen, Freddy Lécué: Interpretable Credit Application Predictions With Counterfactual Explanations. CoRR abs/1811.05245 (2018)



Naive Bayes model

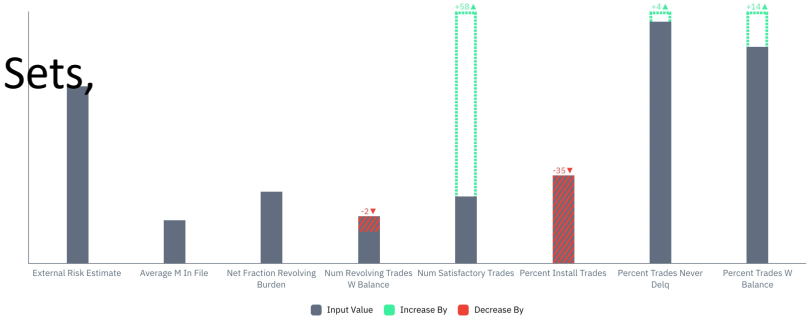
Igor Kononenko. Machine learning for medical diagnosis: history, state of the art and perspective. Artificial Intelligence in Medicine, 23:89–109, 2001.

Overview of explanation in Machine Learning (1)

- All except Artificial Neural Network

Interpretable Models:

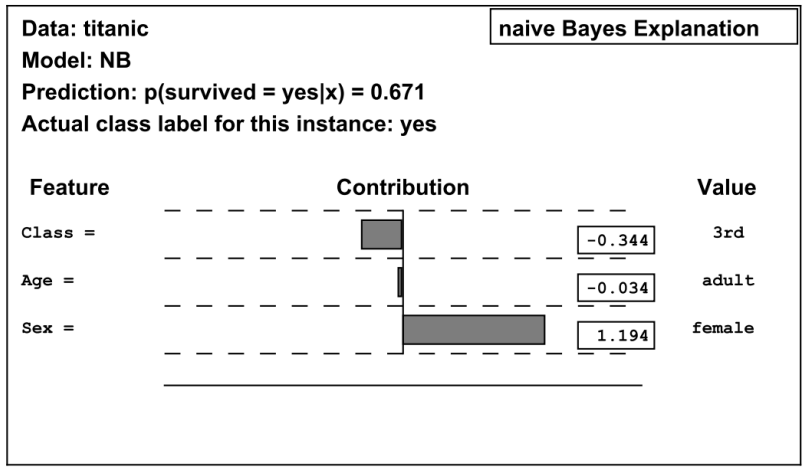
- Decision Trees, Lists and Sets,
- GAMs,
- GLMs,
- Linear regression,
- Logistic regression,
- KNNs



Counterfactual What-if

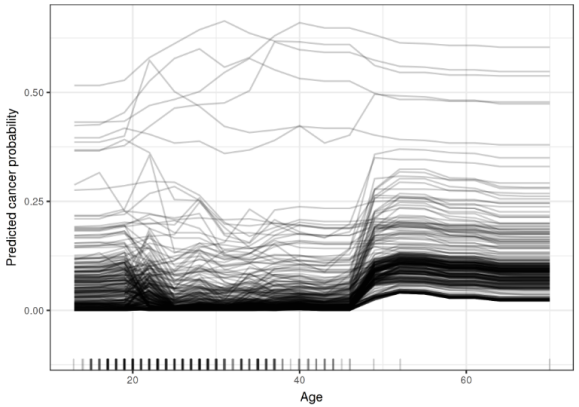
Brent D. Mittelstadt, Chris Russell, Sandra Wachter: Explaining Explanations in AI. FAT 2019: 279-288

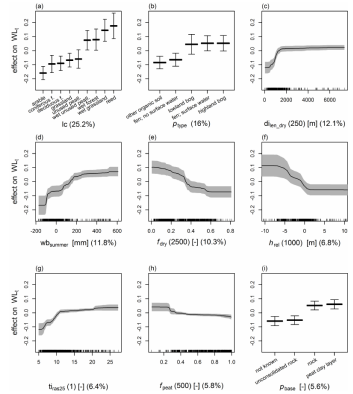
Rory Mc Grath, Luca Costabello, Chan Le Van, Paul Sweeney, Farbod Kamiab, Zhao Shen, Freddy Lécué: Interpretable Credit Application Predictions With Counterfactual Explanations. CoRR abs/1811.05245 (2018)

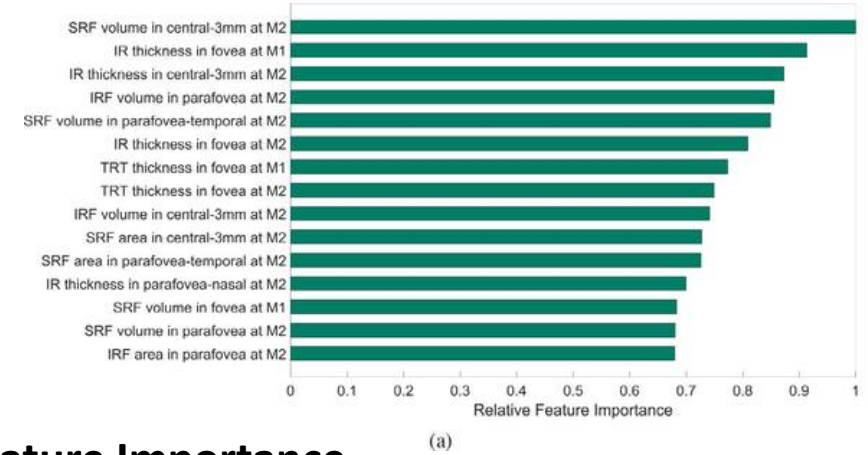


Naive Bayes model

Igor Kononenko. Machine learning for medical diagnosis: history, state of the art and perspective. Artificial Intelligence in Medicine, 23:89–109, 2001.



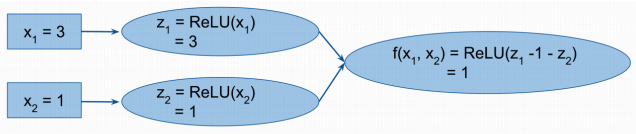




Feature Importance
Partial Dependence Plot
Individual Conditional Expectation
Sensitivity Analysis

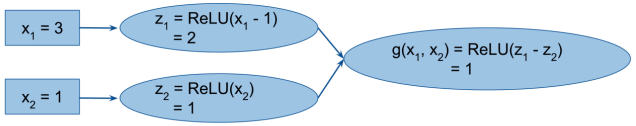
Overview of explanation in Machine Learning (2)

- Only Artificial Neural Network



Network $f(x_1, x_2)$
Attributions at $x_1 = 3, x_2 = 1$

Integrated gradients	$x_1 = 1.5,\ x_2 = -0.5$
DeepLift	$x_1 = 1.5,\ x_2 = -0.5$
LRP	$x_1 = 1.5,\ x_2 = -0.5$



Network $g(x_1, x_2)$
Attributions at $x_1 = 3, x_2 = 1$

Integrated gradients	$x_1 = 1.5,\ x_2 = -0.5$
DeepLift	$x_1 = 2,\ x_2 = -1$
LRP	$x_1 = 2,\ x_2 = -1$

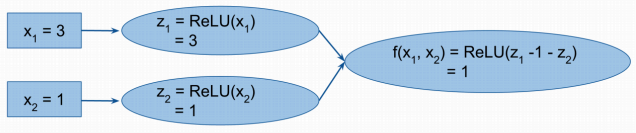
Attribution for Deep Network (Integrated gradient-based)

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In ICML, pp. 3319–3328, 2017.

Avanti Shrikumar, Peyton Greenside, Anshul Kundaje: Learning Important Features Through Propagating Activation Differences. ICML 2017: 3145-3153

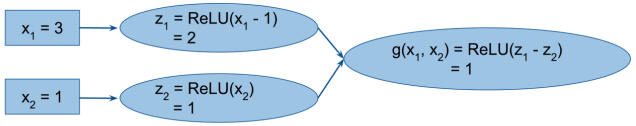
Overview of explanation in Machine Learning (2)

- Only Artificial Neural Network



Network $f(x_1, x_2)$
Attributions at $x_1 = 3, x_2 = 1$

Integrated gradients	$x_1 = 1.5,\ x_2 = -0.5$
DeepLift	$x_1 = 1.5,\ x_2 = -0.5$
LRP	$x_1 = 1.5,\ x_2 = -0.5$



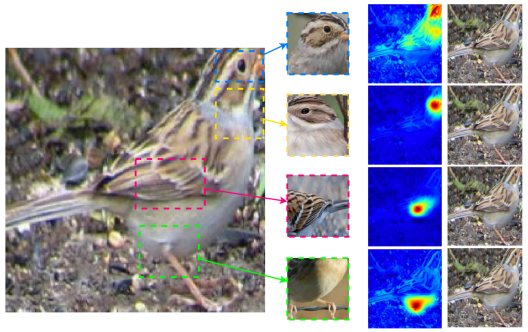
Network $g(x_1, x_2)$
Attributions at $x_1 = 3, x_2 = 1$

Integrated gradients	$x_1 = 1.5,\ x_2 = -0.5$
DeepLift	$x_1 = 2,\ x_2 = -1$
LRP	$x_1 = 2,\ x_2 = -1$

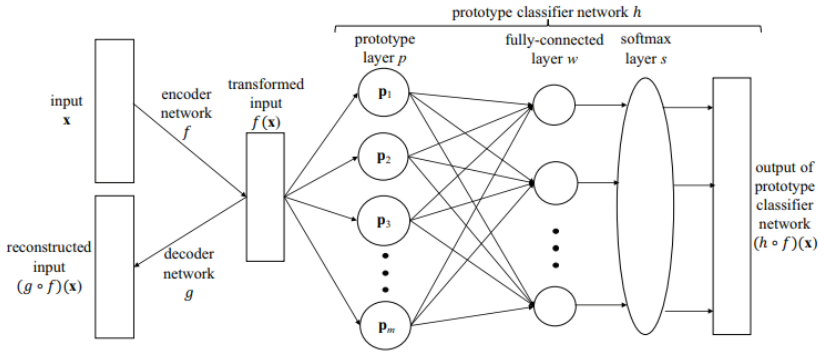
Attribution for Deep Network (Integrated gradient-based)

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In ICML, pp. 3319–3328, 2017.

Avanti Shrikumar, Peyton Greenside, Anshul Kundaje: Learning Important Features Through Propagating Activation Differences. ICML 2017: 3145-3153



Chaofan Chen, Oscar Li, Alina Barnett, Jonathan Su, Cynthia Rudin: This looks like that: deep learning for interpretable image recognition. CoRR abs/1806.10574 (2018)

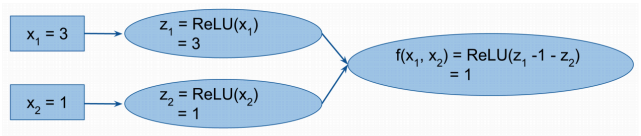


Auto-encoder / Prototype

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

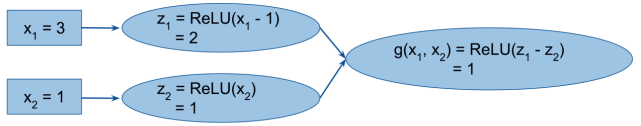
Overview of explanation in Machine Learning (2)

- Only Artificial Neural Network



Network $f(x_1, x_2)$
Attributions at $x_1 = 3, x_2 = 1$

Integrated gradients	$x_1 = 1.5,\ x_2 = -0.5$
DeepLift	$x_1 = 1.5,\ x_2 = -0.5$
LRP	$x_1 = 1.5,\ x_2 = -0.5$



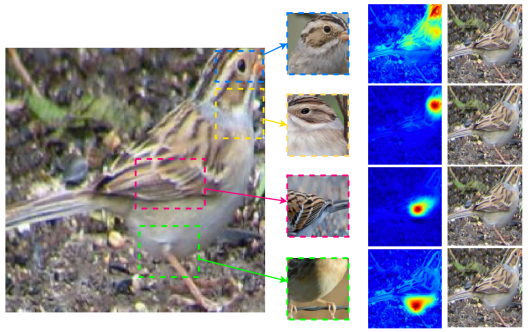
Network $g(x_1, x_2)$
Attributions at $x_1 = 3, x_2 = 1$

Integrated gradients	$x_1 = 1.5,\ x_2 = -0.5$
DeepLift	$x_1 = 2,\ x_2 = -1$
LRP	$x_1 = 2,\ x_2 = -1$

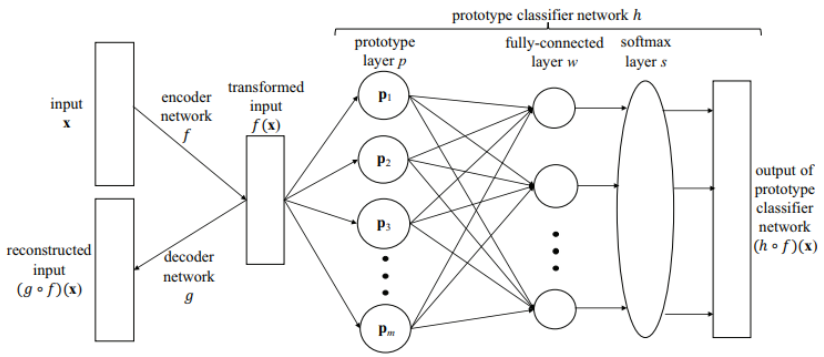
Attribution for Deep Network (Integrated gradient-based)

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In ICML, pp. 3319–3328, 2017.

Avanti Shrikumar, Peyton Greenside, Anshul Kundaje: Learning Important Features Through Propagating Activation Differences. ICML 2017: 3145-3153

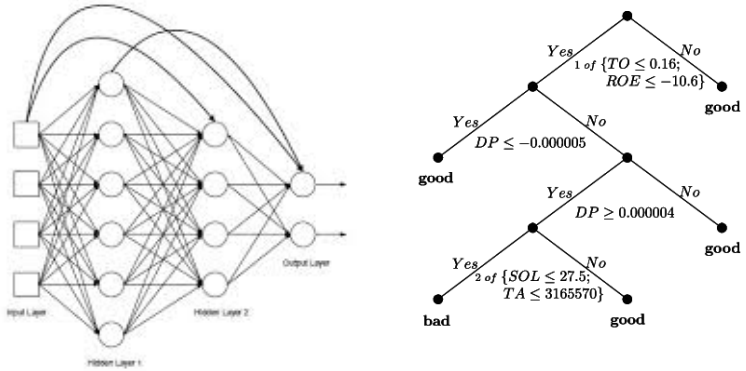


Chaofan Chen, Oscar Li, Alina Barnett, Jonathan Su, Cynthia Rudin: This looks like that: deep learning for interpretable image recognition. CoRR abs/1806.10574 (2018)



Auto-encoder / Prototype

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

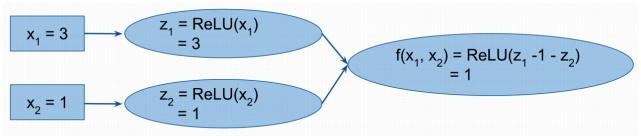


Surogate Model

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

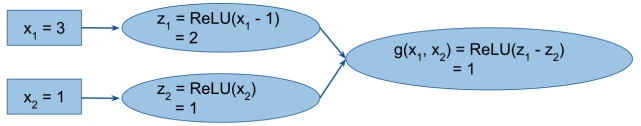
Overview of explanation in Machine Learning (2)

- Only Artificial Neural Network



Network $f(x_1, x_2)$
Attributions at $x_1 = 3, x_2 = 1$

Integrated gradients	$x_1 = 1.5,\ x_2 = -0.5$
DeepLift	$x_1 = 1.5,\ x_2 = -0.5$
LRP	$x_1 = 1.5,\ x_2 = -0.5$



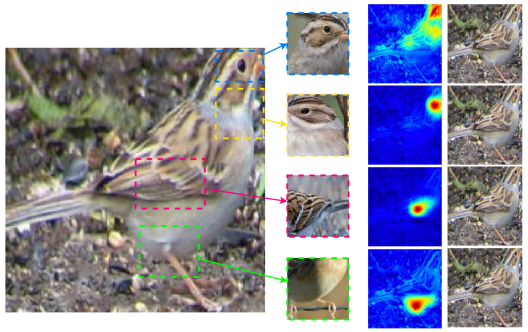
Network $g(x_1, x_2)$
Attributions at $x_1 = 3, x_2 = 1$

Integrated gradients	$x_1 = 1.5,\ x_2 = -0.5$
DeepLift	$x_1 = 2,\ x_2 = -1$
LRP	$x_1 = 2,\ x_2 = -1$

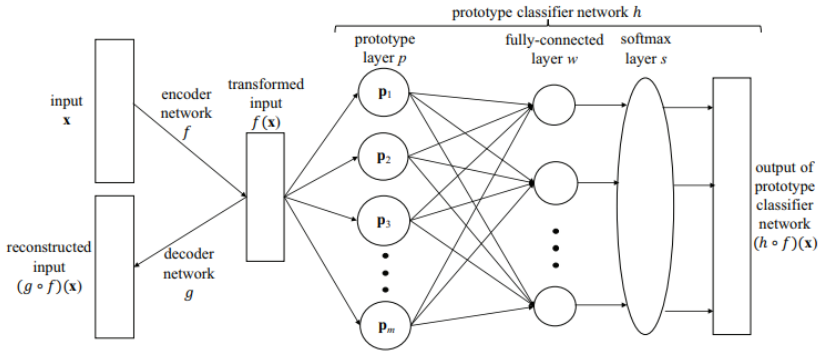
Attribution for Deep Network (Integrated gradient-based)

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In ICML, pp. 3319–3328, 2017.

Avanti Shrikumar, Peyton Greenside, Anshul Kundaje: Learning Important Features Through Propagating Activation Differences. ICML 2017: 3145-3153

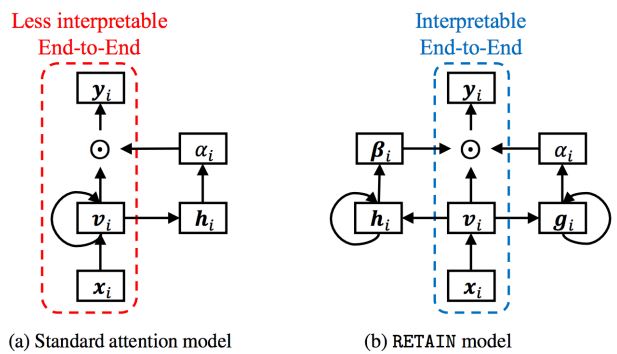


Chaofan Chen, Oscar Li, Alina Barnett, Jonathan Su, Cynthia Rudin: This looks like that: deep learning for interpretable image recognition. CoRR abs/1806.10574 (2018)



Auto-encoder / Prototype

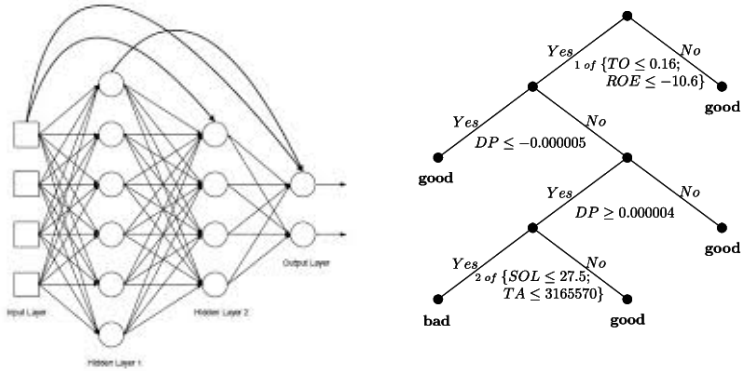
Oscar Li, Hao Liu, Chaofan Chen, Cynthia Rudin: Deep Learning for Case-Based Reasoning Through Prototypes: A Neural Network That Explains Its Predictions. AAAI 2018: 3530-3537



Attention Mechanism

Edward Choi, Mohammad Taha Bahadori, Jimeng Sun, Joshua Kulas, Andy Schuetz, Walter F. Stewart: RETAIN: An Interpretable Predictive Model for Healthcare using Reverse Time Attention Mechanism. NIPS 2016: 3504-3512

D. Bahdanau, K. Cho, and Y. Bengio. Neural machine translation by jointly learning to align and translate. International Conference on Learning Representations, 2015

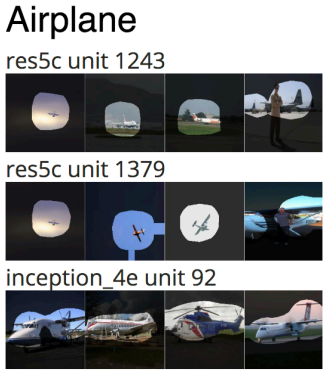


Surogate Model

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

Overview of explanation in Machine Learning (3)

- Computer Vision

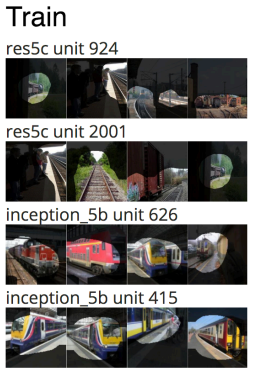


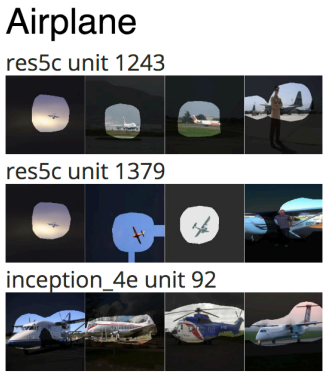
Interpretable Units

David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, Antonio Torralba: Network Dissection: Quantifying Interpretability of Deep Visual Representations. CVPR 2017: 3319-3327

Overview of explanation in Machine Learning (3)

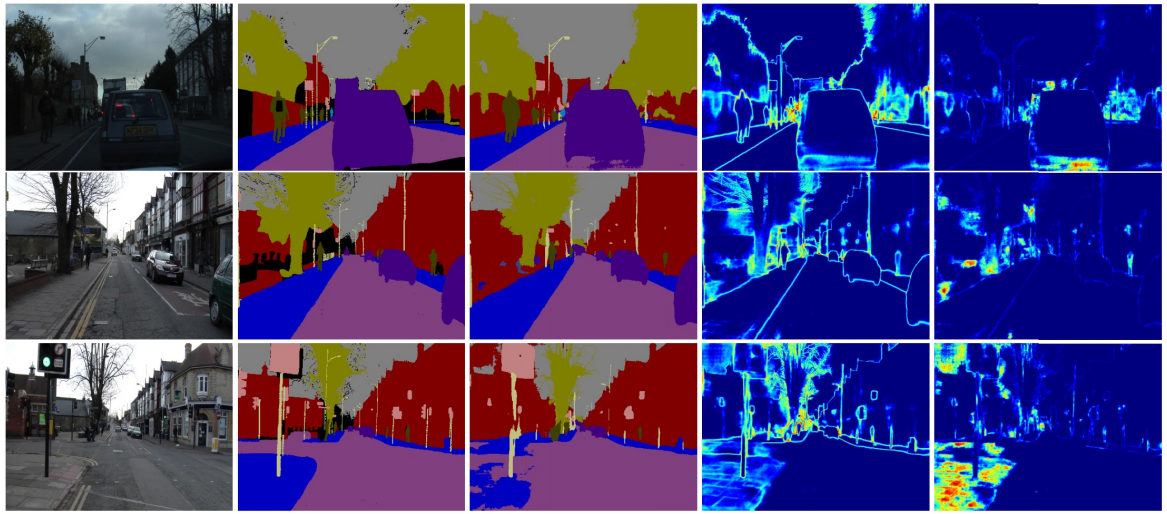
- Computer Vision





Interpretable Units

David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, Antonio Torralba: Network Dissection: Quantifying Interpretability of Deep Visual Representations. CVPR 2017: 3319-3327



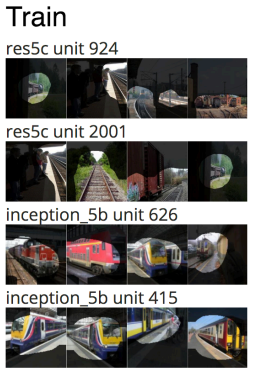
(a) Input Image (b) Ground Truth (c) Semantic Segmentation (d) Aleatoric Uncertainty (e) Epistemic Uncertainty

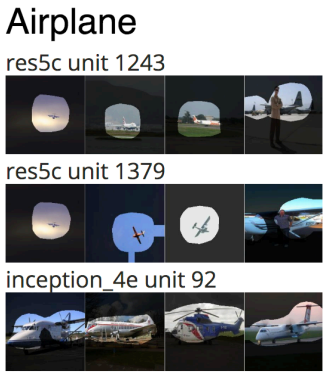
Uncertainty Map

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

Overview of explanation in Machine Learning (3)

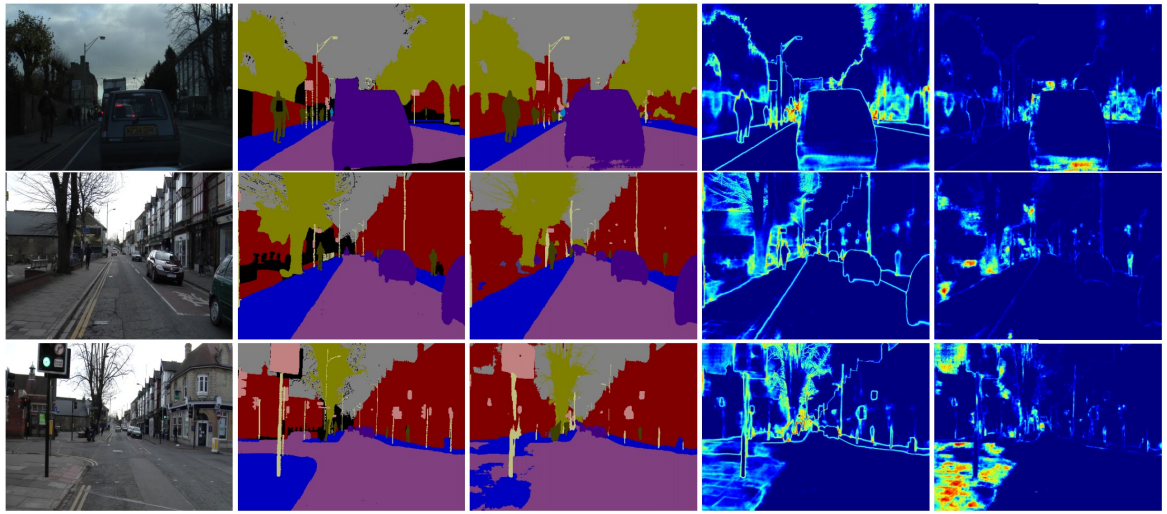
- Computer Vision





Interpretable Units

David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, Antonio Torralba: Network Dissection: Quantifying Interpretability of Deep Visual Representations. CVPR 2017: 3319-3327



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

Uncertainty Map

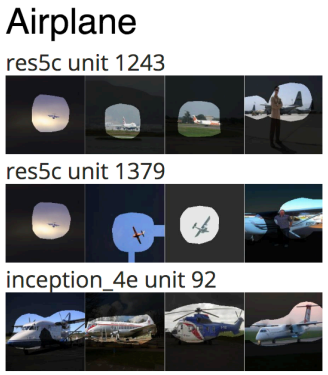
Alex Kendall, Yarin Gal: What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision? NIPS 2017: 5580-5590

Visual Explanation

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

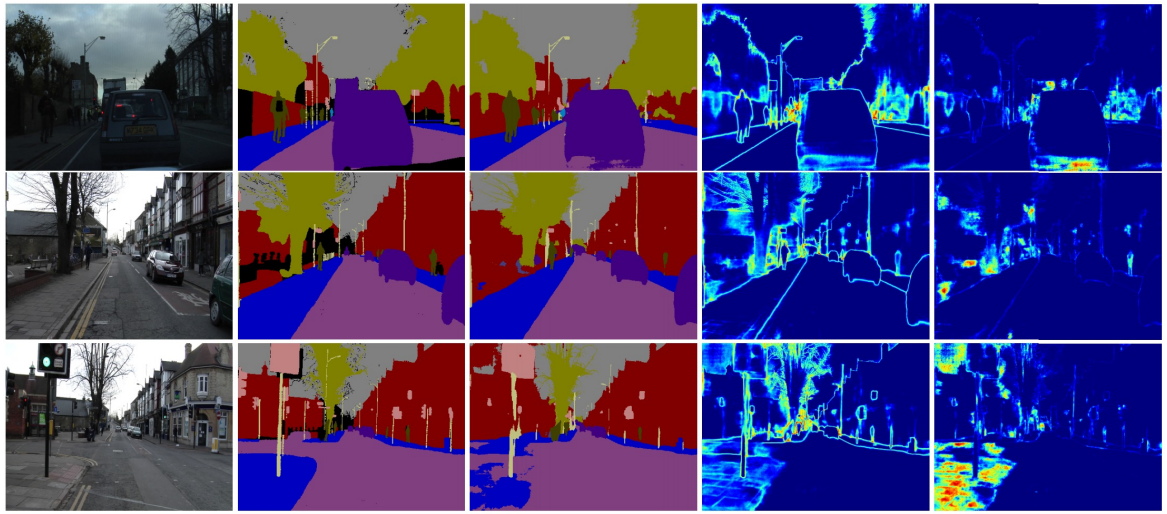
Overview of explanation in Machine Learning (3)

- Computer Vision



Interpretable Units

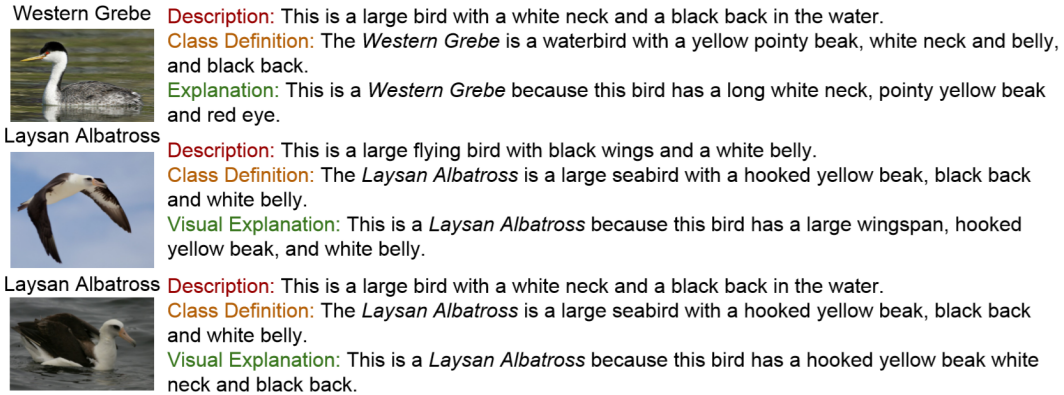
David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, Antonio Torralba: Network Dissection: Quantifying Interpretability of Deep Visual Representations. CVPR 2017: 3319-3327



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

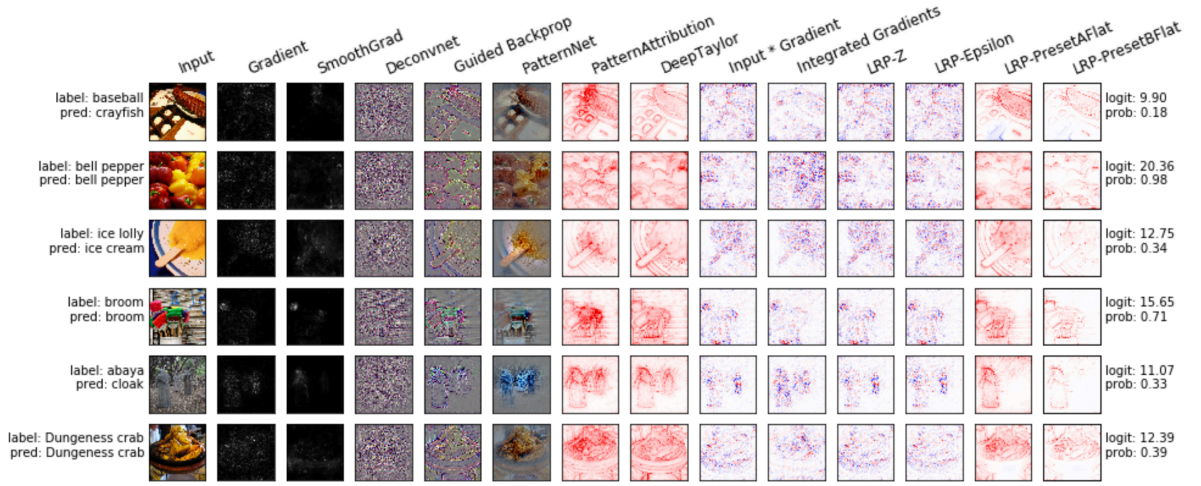
Uncertainty Map

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



Visual Explanation

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



Saliency Map

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

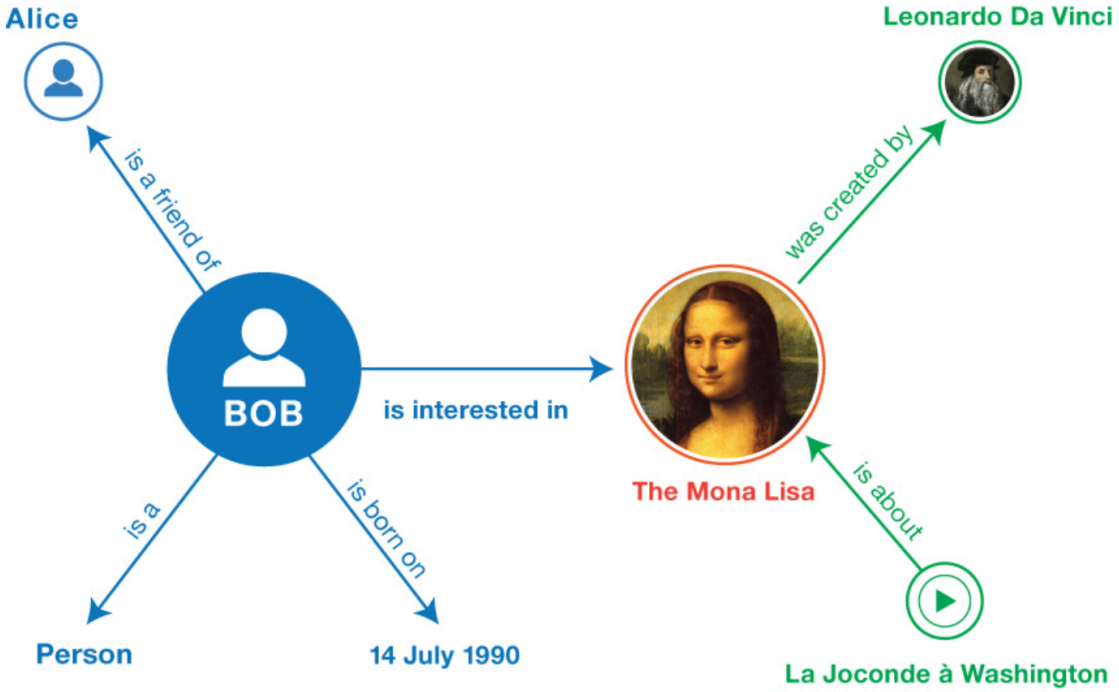
On the Role of Knowledge Graphs in Explainable Machine Learning

On the Role of Knowledge Graph in Explainable AI - under open review at the Semantic Web Journal - http://www.semantic-web-journal.net/content/role-knowledge-graphs-explainable-ai

Knowledge Graph (1)

- 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*



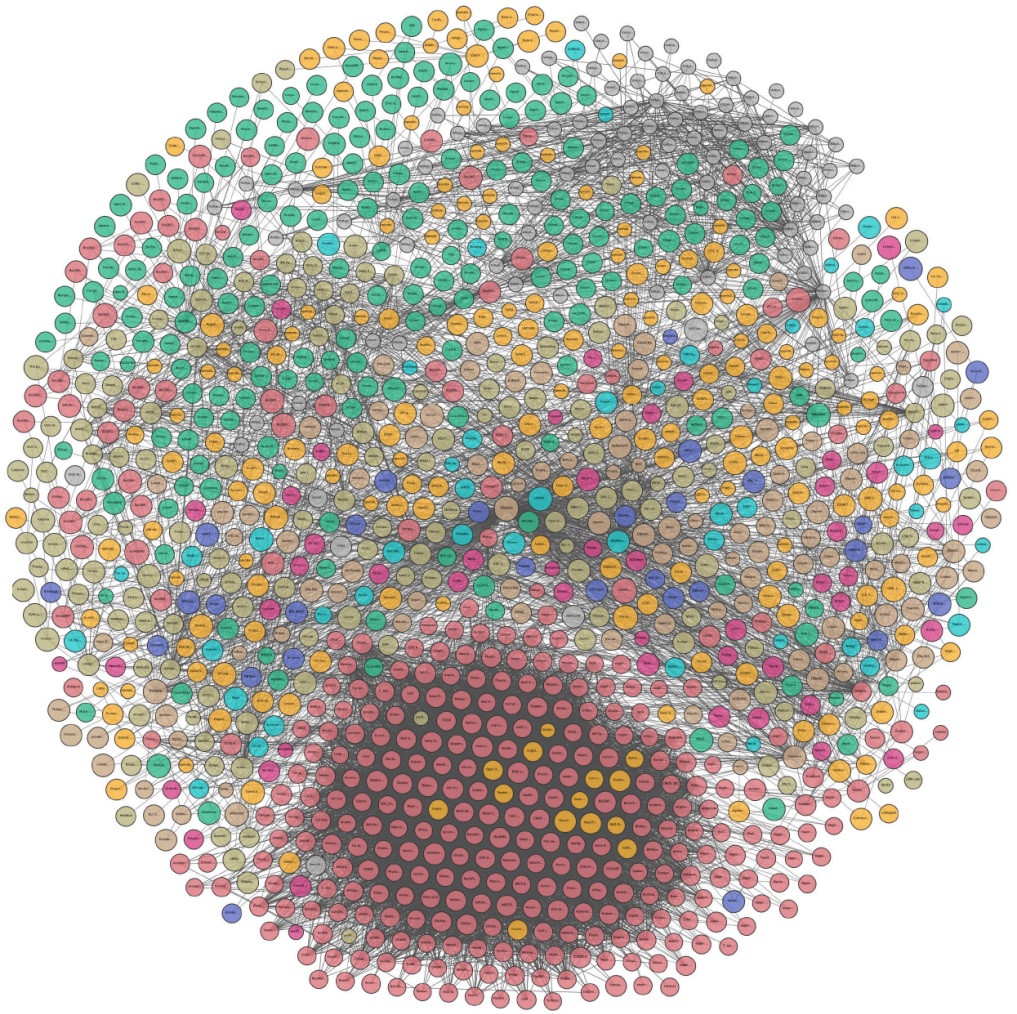
Knowledge Graph (2)

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 Graph 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:

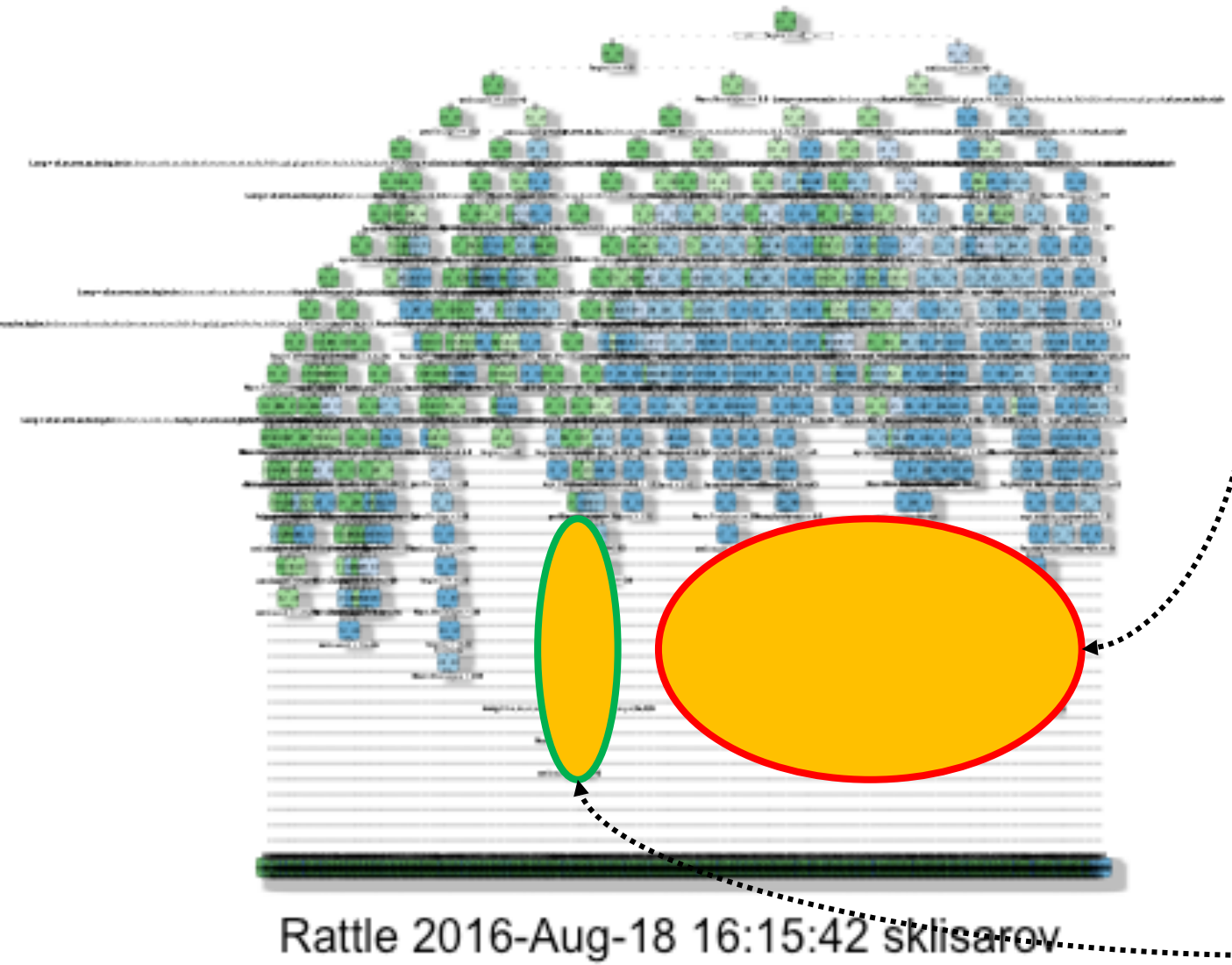
- **Freebase** (40M entities) - 71% of persons without a birthplace, 75% without a nationality, even worse for other relation types [Dong et al. 2014]
- **DBpedia** (20M entities) - 61% of persons without a birthplace, 58% of scientists missing why they are popular [Krompaß et al. 2015]

Relational Learning can help us overcoming these issues.

Knowledge Graph Embeddings in Machine Learning

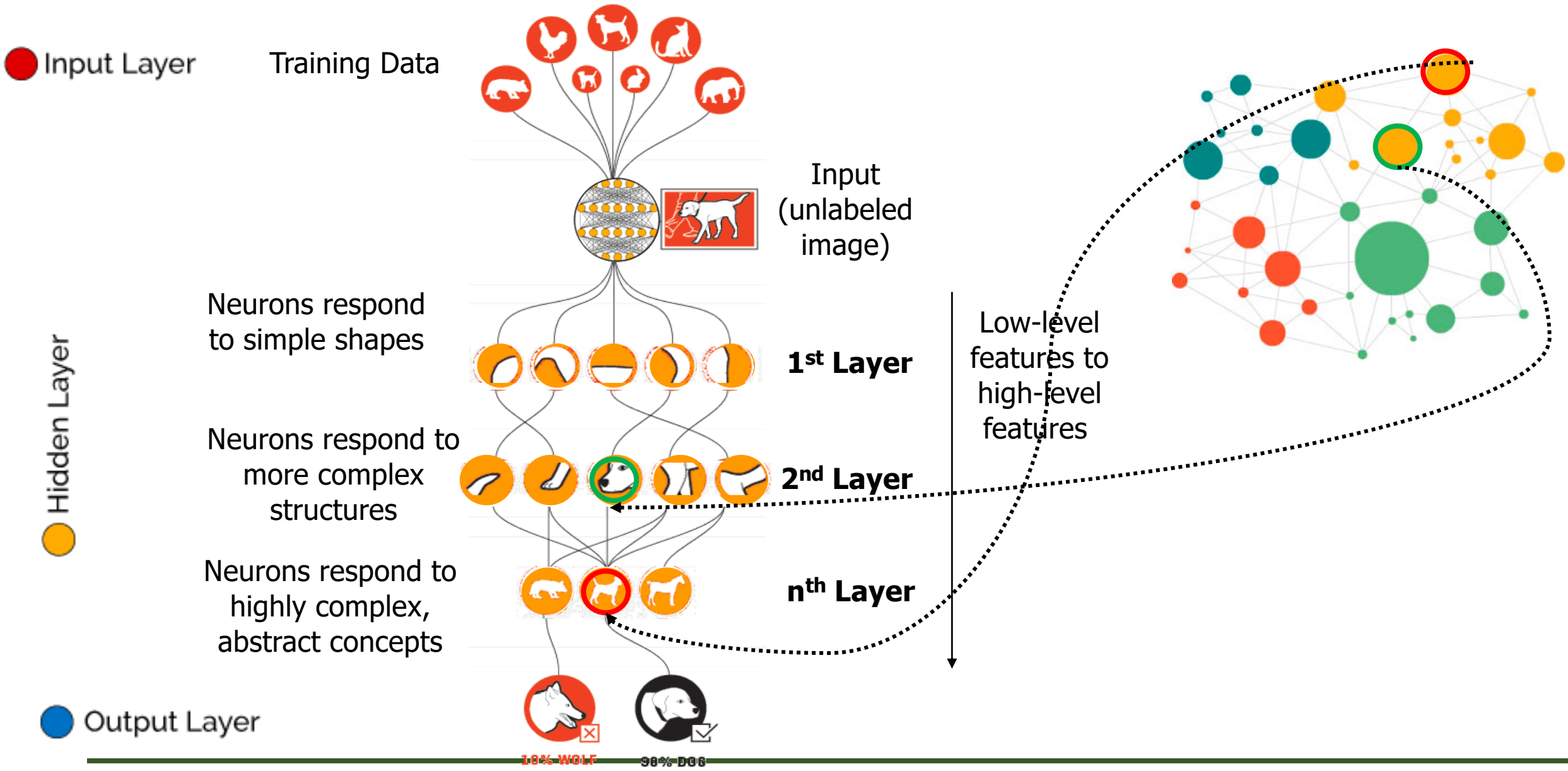
https://stats.stackexchange.com/questions/230581/decision-tree-too-large-to-interpret

Knowledge Graph for Decision Trees

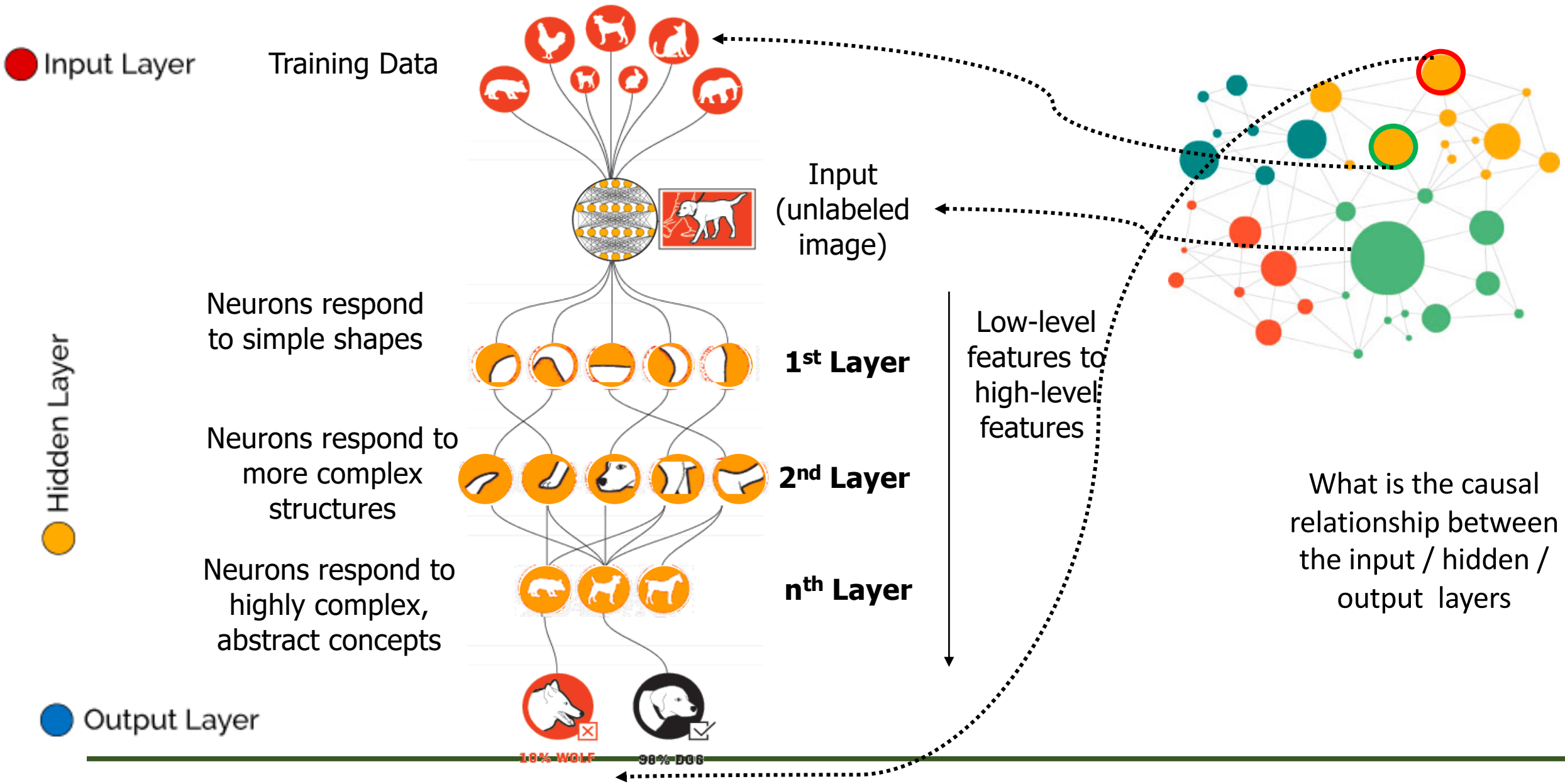


https://stats.stackexchange.com/questions/230581/decision-tree-too-large-to-interpret

Knowledge Graph for Deep Neural Network (1)



Knowledge Graph for Deep Neural Network (2)



Knowledge Graph for Personalized XAI

Description 1: This is an orange train accident

Description 2: This is an train accident between two speed merchant trains of characteristics X43-B and Y33-C in a dry environment

Description 3: This is a public transportation accident

Knowledge Graph for Explaining Transfer Learning

"How to explain transfer learning with appropriate knowledge representation?

Proceedings of the Sixteenth International Conference on Principles of Knowledge Representation and Reasoning (KR 2018)

Knowledge-Based Transfer Learning Explanation

Jiaoyan Chen
Department of Computer Science
University of Oxford, UK

Freddy Lecue
INRIA, France
Accenture Labs, Ireland

Jeff Z. Pan
Department of Computer Science
University of Aberdeen, UK

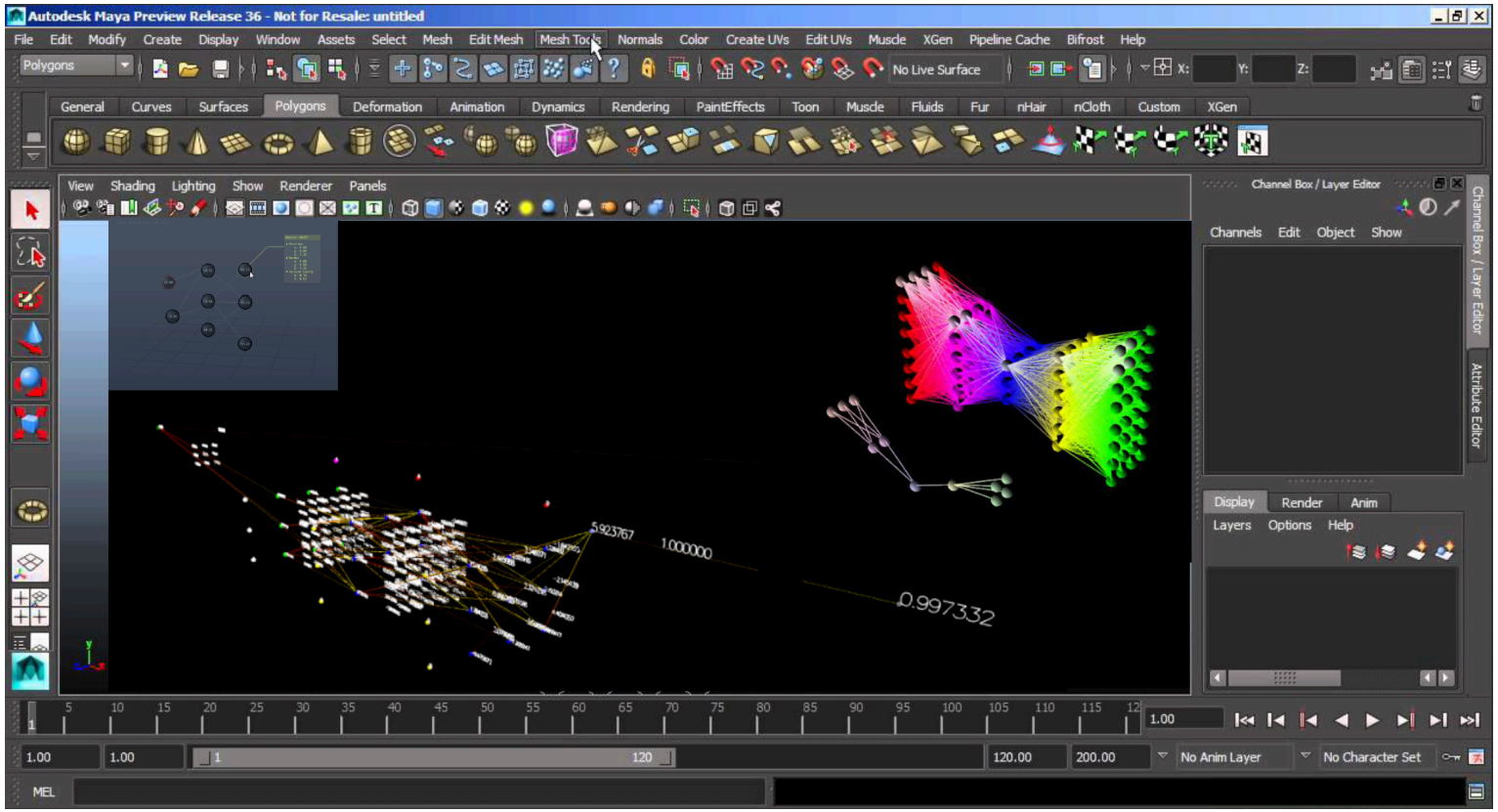
Ian Horrocks
Department of Computer Science
University of Oxford, UK

Huajun Chen
College of Computer Science, Zhejiang University, China
Alibaba-Zhejian University Frontier Technology Research Center

On the Role of Reasoning in Explainable Machine Learning

Applications

Debugging Artificial Neural Networks – Industry Agnostic



Challenge: Designing Artificial Neural Network architectures requires lots of experimentation (i.e., training phases) and parameters tuning (optimization strategy, learning rate, number of layers…) to reach optimal and robust machine learning models.

AI Technology: Artificial Neural Network

XAI Technology: Artificial Neural Network, 3D Modeling and Simulation Platform For AI

Zetane.com

Explaining Visual Question Answering – Industry Agnostic

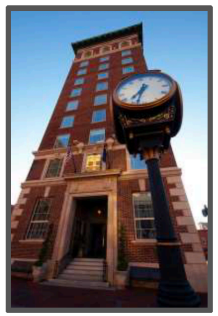
Tabular QA

Rank	Nation	Gold	Silver	Bronze	Total
1	India	102	58	37	197
2	Nepal	32	10	24	65
3	Sri Lanka	16	42	62	120
4	Pakistan	10	36	30	76
5	Bangladesh	2	10	35	47
6	Bhutan	1	6	7	14
7	Maldives	0	0	4	4

Q: How many medals did India win?
A: 197

Neural Programmer (2017) model
33.5% accuracy on WikiTableQuestions

Visual QA



Q: How symmetrical are the white bricks on either side of the building?
A: very

Kazemi and Elqursh (2017) model.
61.1% on VQA 1.0 dataset
(state of the art = 66.7%)

Reading Comprehension

Peyton Manning became the first quarterback ever to lead two different teams to multiple Super Bowls. He is also the oldest quarterback ever to play in a Super Bowl at age 39. The past record was held by John Elway, who led the Broncos to victory in Super Bowl XXXIII at age 38 and is currently Denver's Executive Vice President of Football Operations and General Manager

Q: Name of the quarterback who was 38 in Super Bowl XXXIII?
A: John Elway

Yu et al (2018) model.
84.6 F-1 score on SQuAD (state of the art)

Challenge: What is the robustness of Visual Question Answering models? What is the impact of semantics?

AI Technology: Artificial Neural Networks.

XAI Technology: Integrated Gradients

Google AI

Q: How symmetrical are the white bricks on either side of the building?
A: very

Q: How asymmetrical are the white bricks on either side of the building?
A: very

Q: How big are the white bricks on either side of the building?
A: very

Q: How fast are the bricks speaking on either side of the building?
A: very

What is the **man** doing? → What is the **tweet** doing?
How many **children** are there? → How many **tweet** are there?

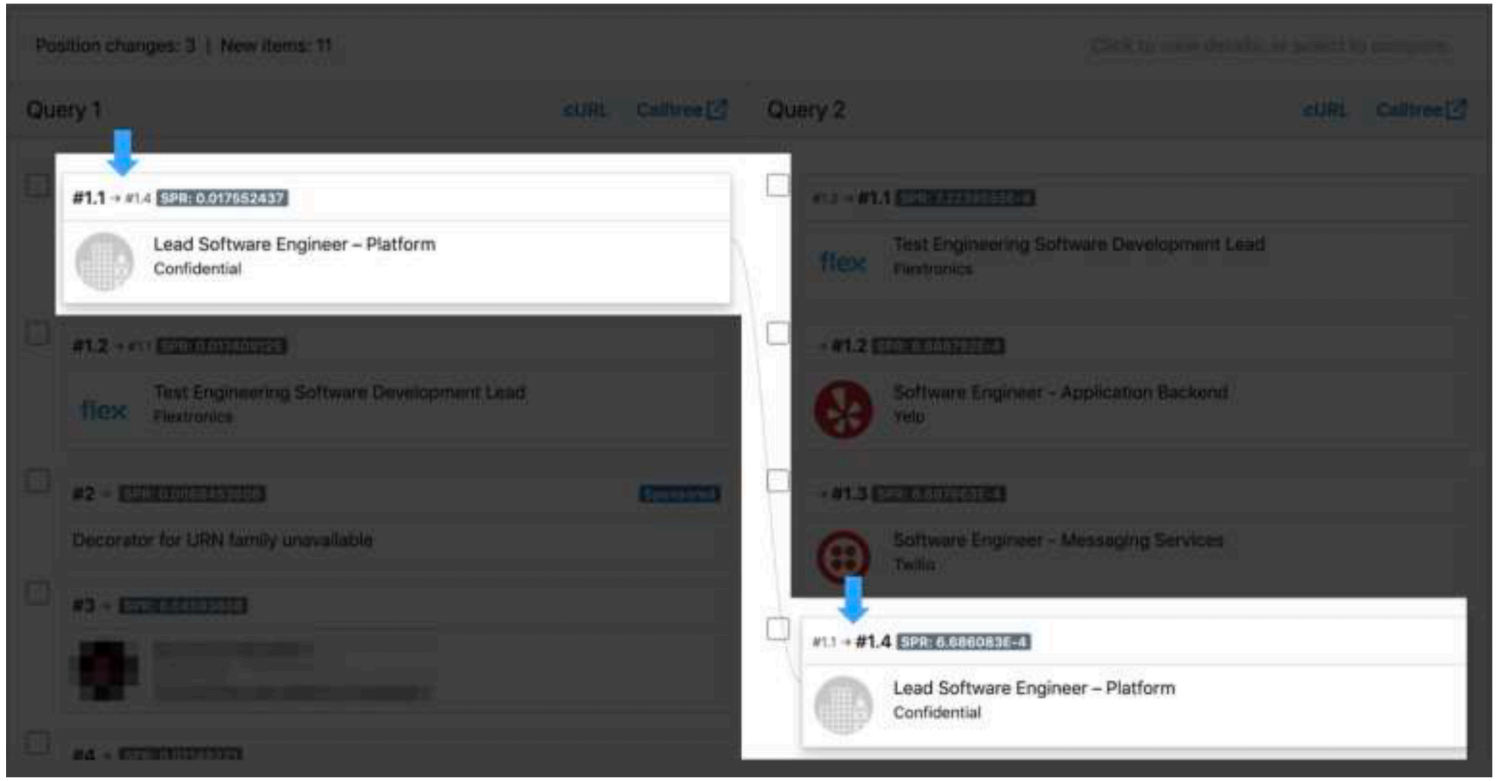
VQA model's response remains the same 75.6
of the time on questions that it originally
answered correctly

Relevance Debugging and Explaining – Industry Agnostic

Challenge: A Machine Learning system can fail in many different points e.g., data features selection, construction, inconsistencies. How to debug bad performance in machine learning models and prediction?

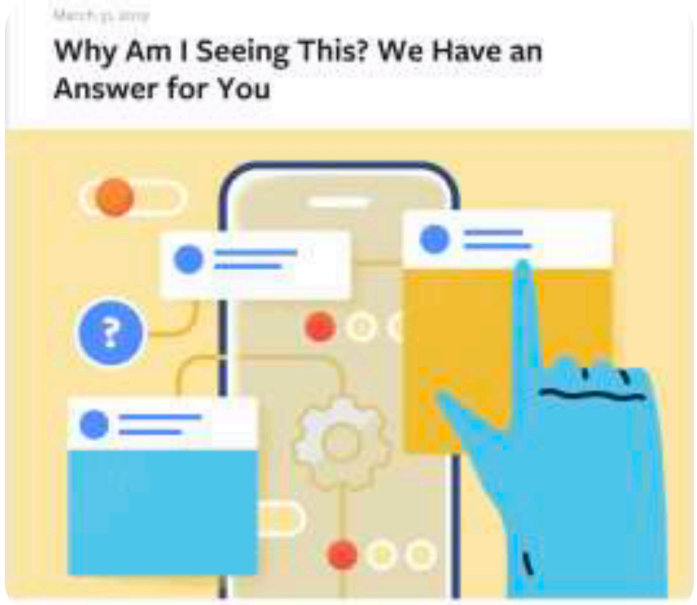
AI Technology: Artificial Neural Networks.

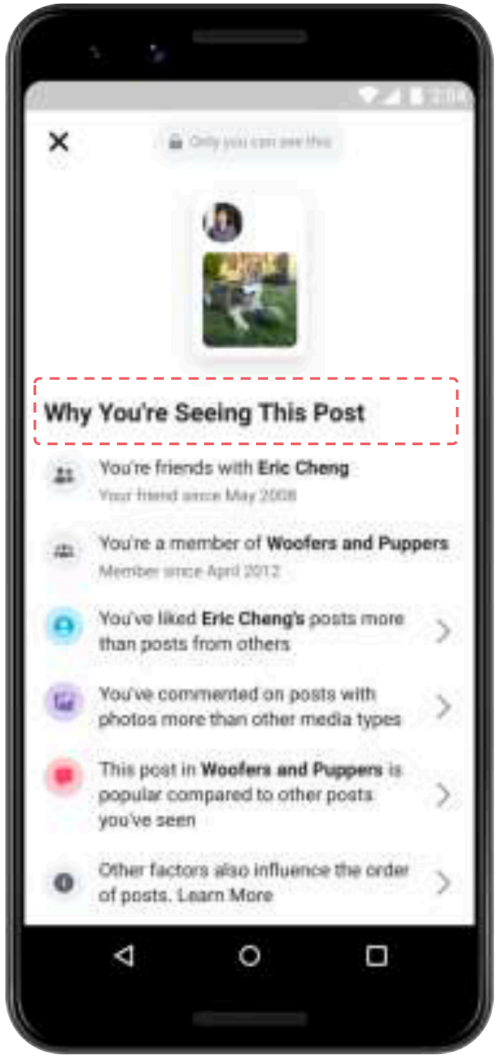
XAI Technology: Model / Prediction comparison



Source: Explainable AI in Industry. KDD 2019 Tutorial. Daniel Qiu, Yucheng Qian

Explaining Recommendation– Social Media



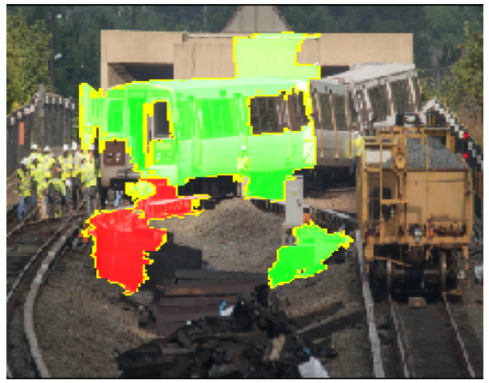


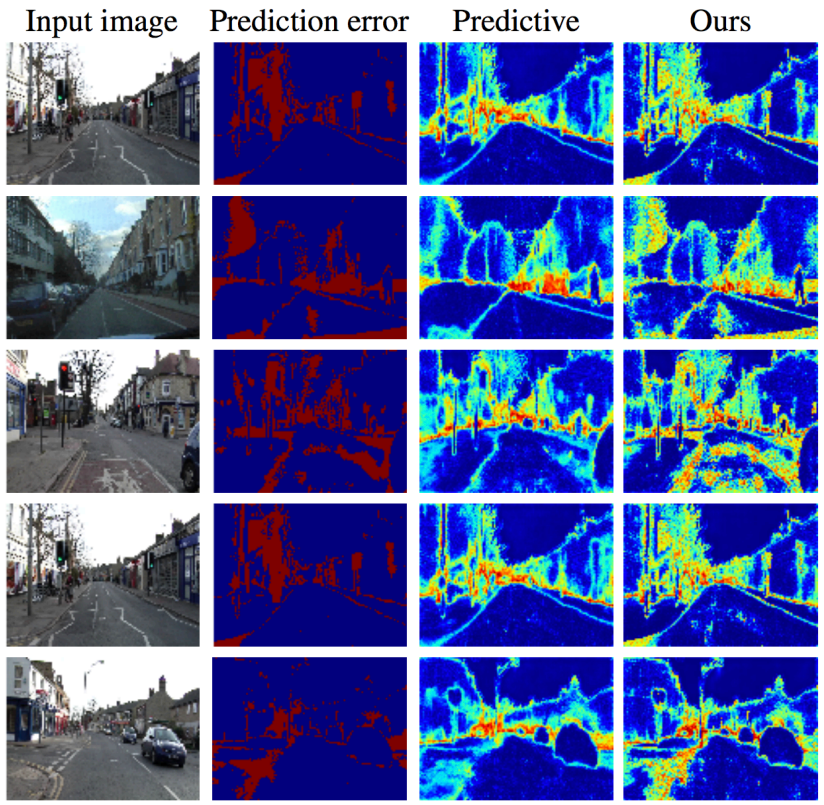
Challenge: How to establish trust between Social Media and their users? Explaining post / news recommendation is crucial for users to engage with content providers.

AI Technology: Artificial Neural Networks.

XAI Technology: Recommendation-based

Obstacle Identification Certification (Trust) - Transportation





THALES

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

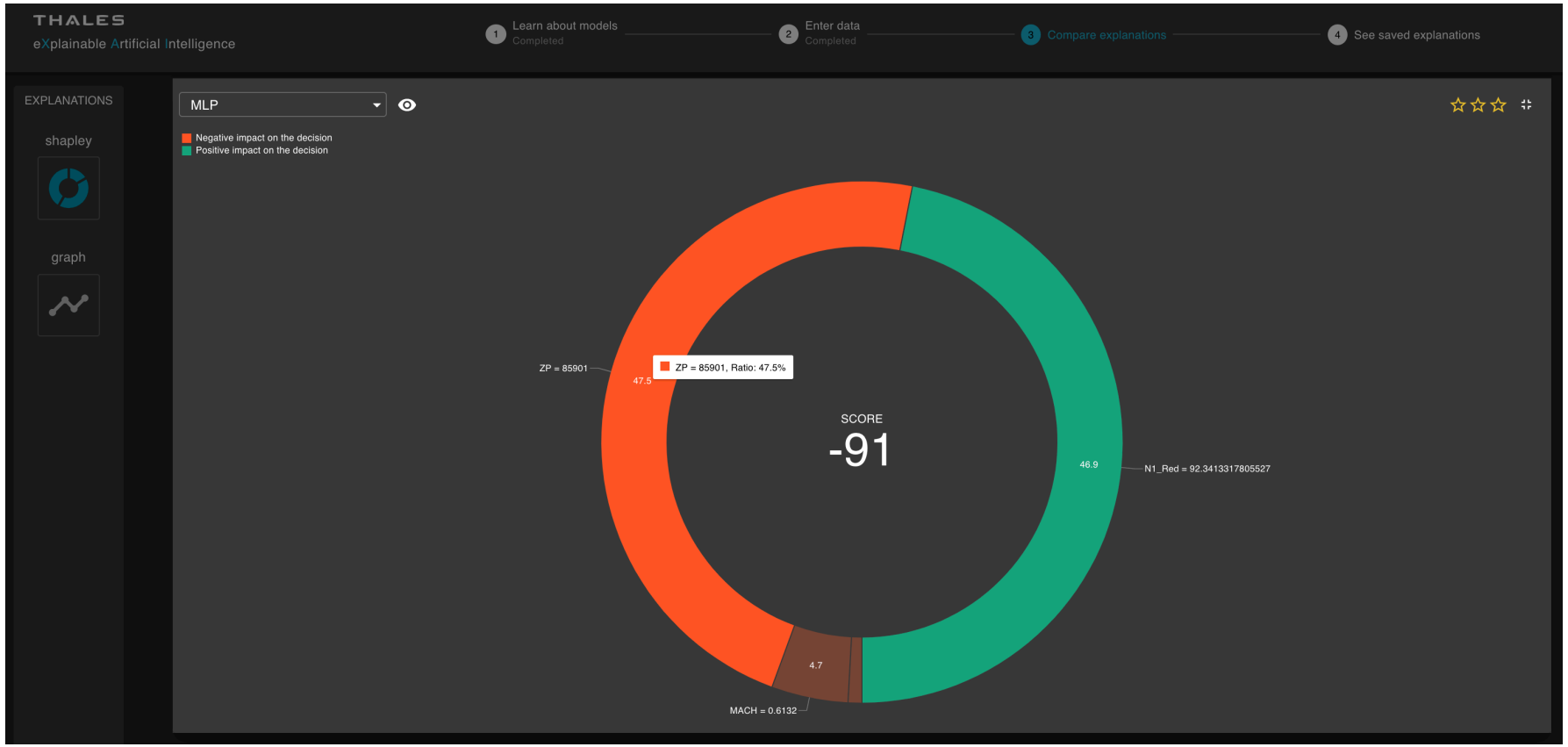
Explaining Flight Performance- Transportation

Challenge: Predicting and explaining aircraft engine performance

AI Technology: Artificial Neural Networks

XAI Technology: Shapely Values

THALES



Explainable On-Time Performance - Transportation

KLM / Transavia Flight Delay Prediction

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

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

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

XAI Technology: Knowledge graph embedded Sequence Learning using LSTMs

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

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

Model Explanation for Sales Prediction - Sales

① What are top driver features for a certain company to have high/low probability to upsell/churn?

① Feature Contributor

② Which top driver features can be perturbed if we want to increase/decrease probability for a certain company?

② Feature Influencer

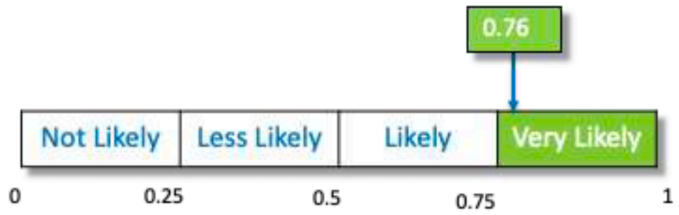
Challenge: How to predict and explain upsell / churn for a company?

AI Technology: Artificial Neural Networks.

XAI Technology: Features importance (contribution, influence), LIME.

Company: CompanyX

Upsell LCP (LinkedIn Career Page)



Top Feature Contributor

- f1: 430.5
- f2: 216
- f3: 10097.57
- f4: 15

Top Feature Influencer (Positive)

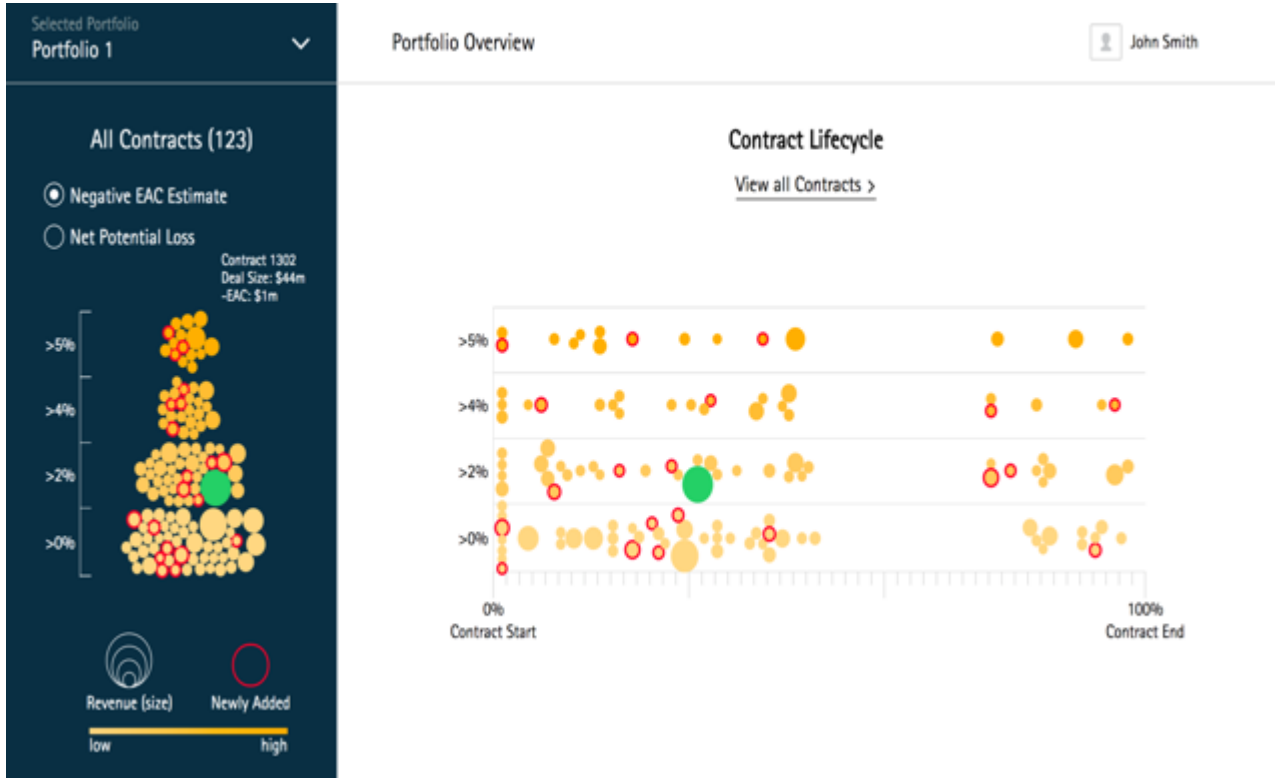
- f5: 0 → 5.4, ↑ 0.03
- f6: 168 → 0, ↑ 0.03
- f7: 0 → 0.24, ↑ 0.02

Top Feature Influencer (Negative)

- f1: 430.5 → 148.7, ↓ 0.20
- f2: 216 → 0, ↓ 0.17
- f8: 423 → 146.0, ↓ 0.07

Source: Explainable AI in Industry. KDD 2019 Tutorial. Jilei Yang, Wei Di, Songtao Guo

Explainable Risk Management - Finance



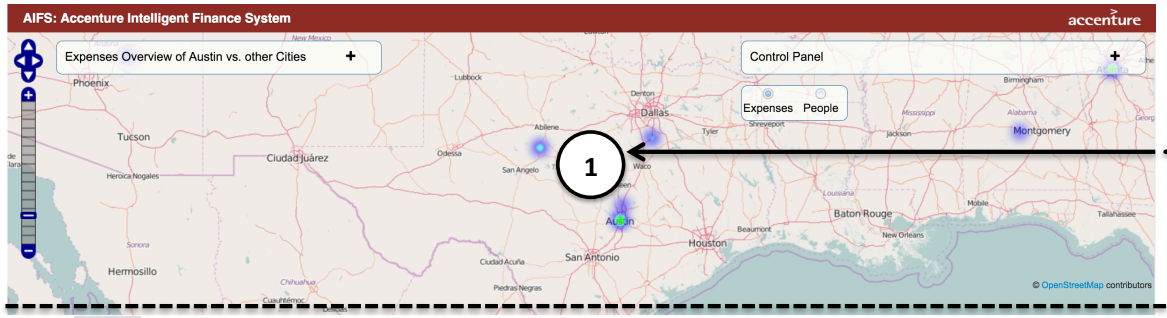
Jiewen Wu, Freddy Lécué, Christophe Guéret, Jer Hayes, Sara van de Moosdijk, Gemma Gallagher, Peter McCanney, Eugene Eichelberger: Personalizing Actions in Context for Risk Management Using Semantic Web Technologies. International Semantic Web Conference (2) 2017: 367-383

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.

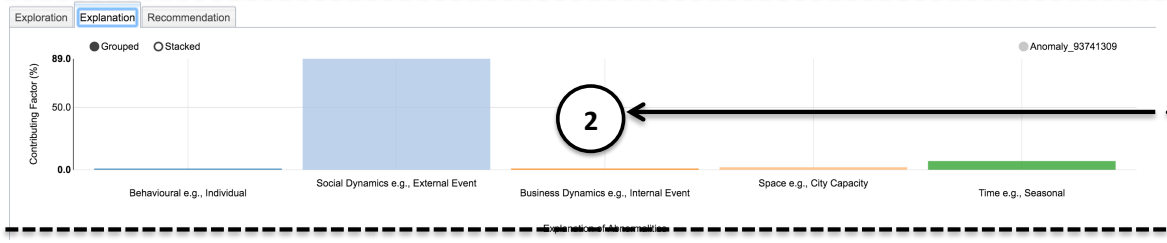
AI Technology: Integration of AI technologies i.e., Machine Learning, Reasoning, Natural Language Processing for building a robust model which can (1) predict revenue loss, (2) recommend corrective actions, and (3) explain why such actions might have a positive impact.

XAI Technology: Knowledge graph embedded Random Forrest

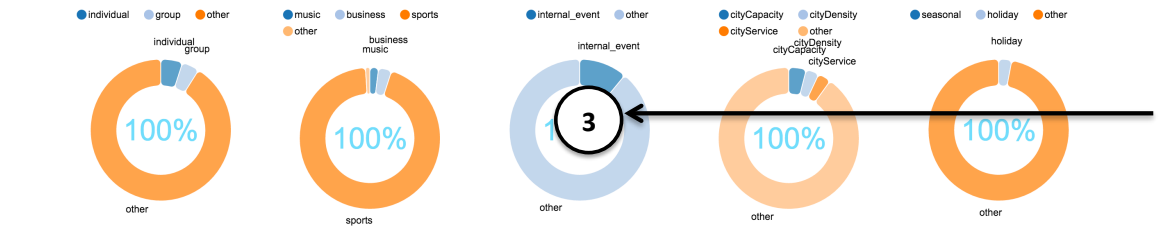
Explainable Anomaly Detection – Finance (Compliance)



Data analysis for spatial interpretation of abnormalities: abnormal expenses

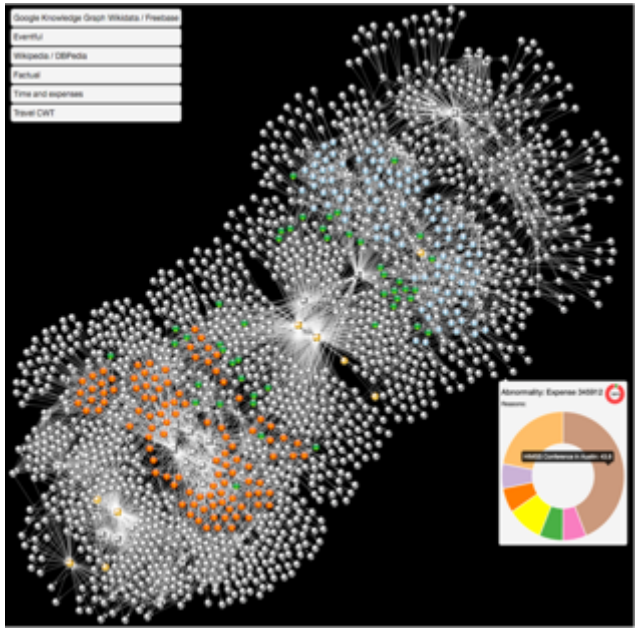


Semantic explanation (structured in classes: fraud, events, seasonal) of abnormalities



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

INNOVATION ARCHITECTURE:
ACCENTURE LABS



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 (1) - Finance

- 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

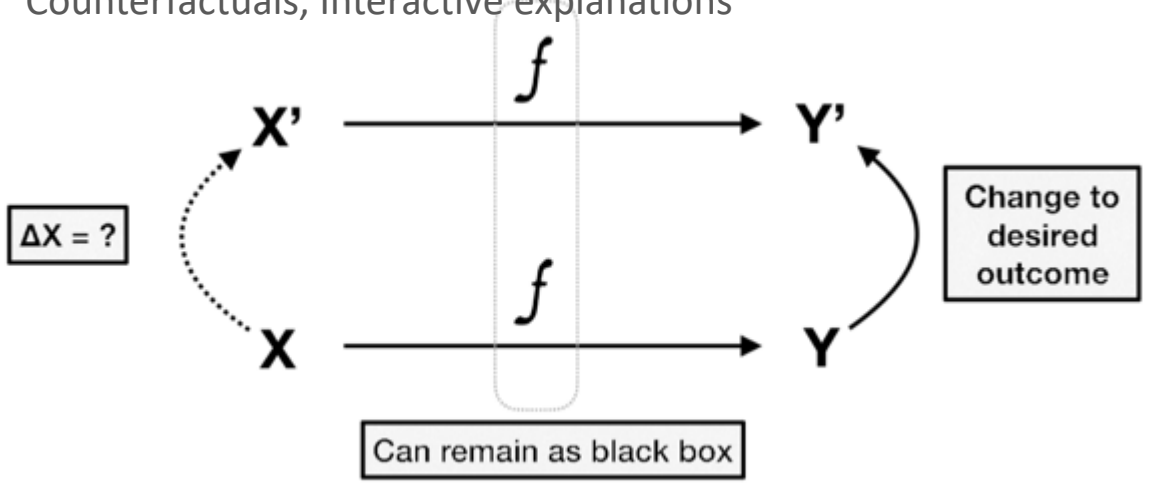
THALES

INNOVATION ARCHITECTURE: ACCENTURE LABS

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.

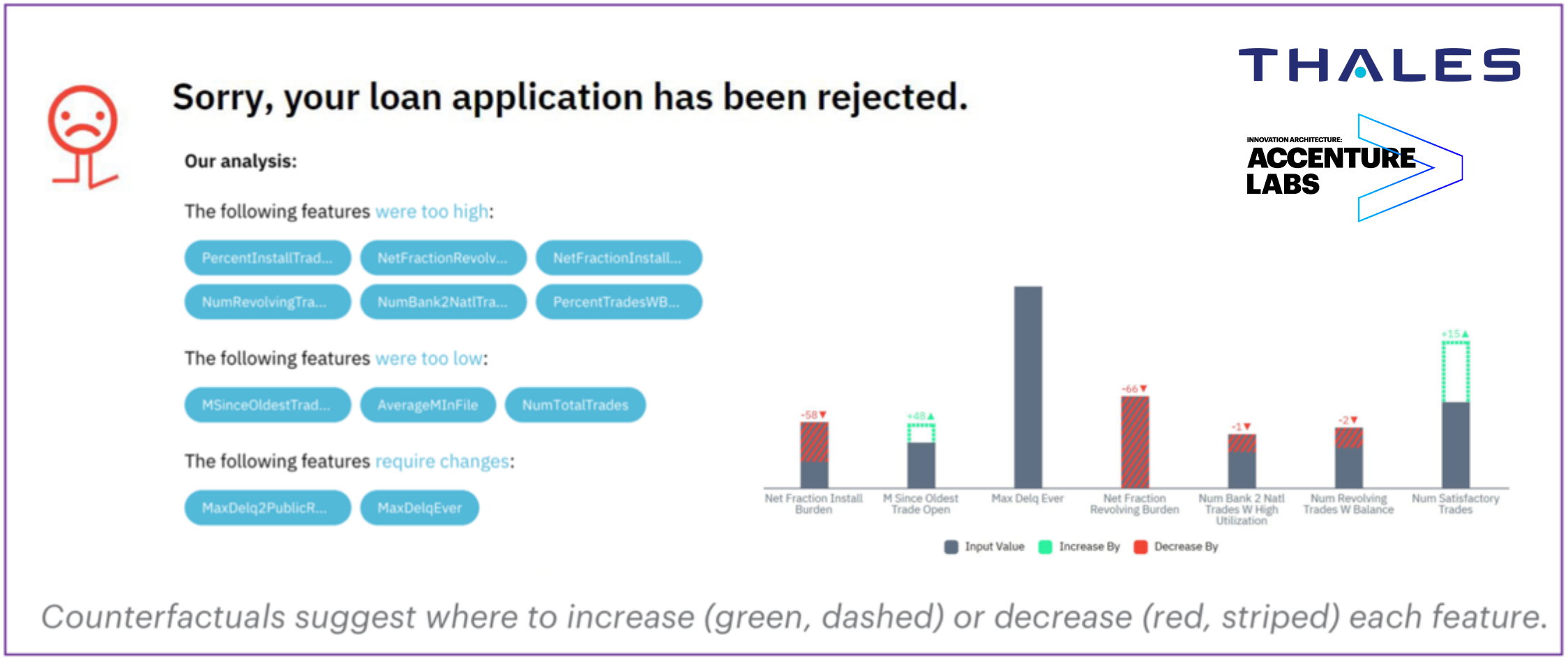
AI Technology: Supervised learning, binary classification.

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



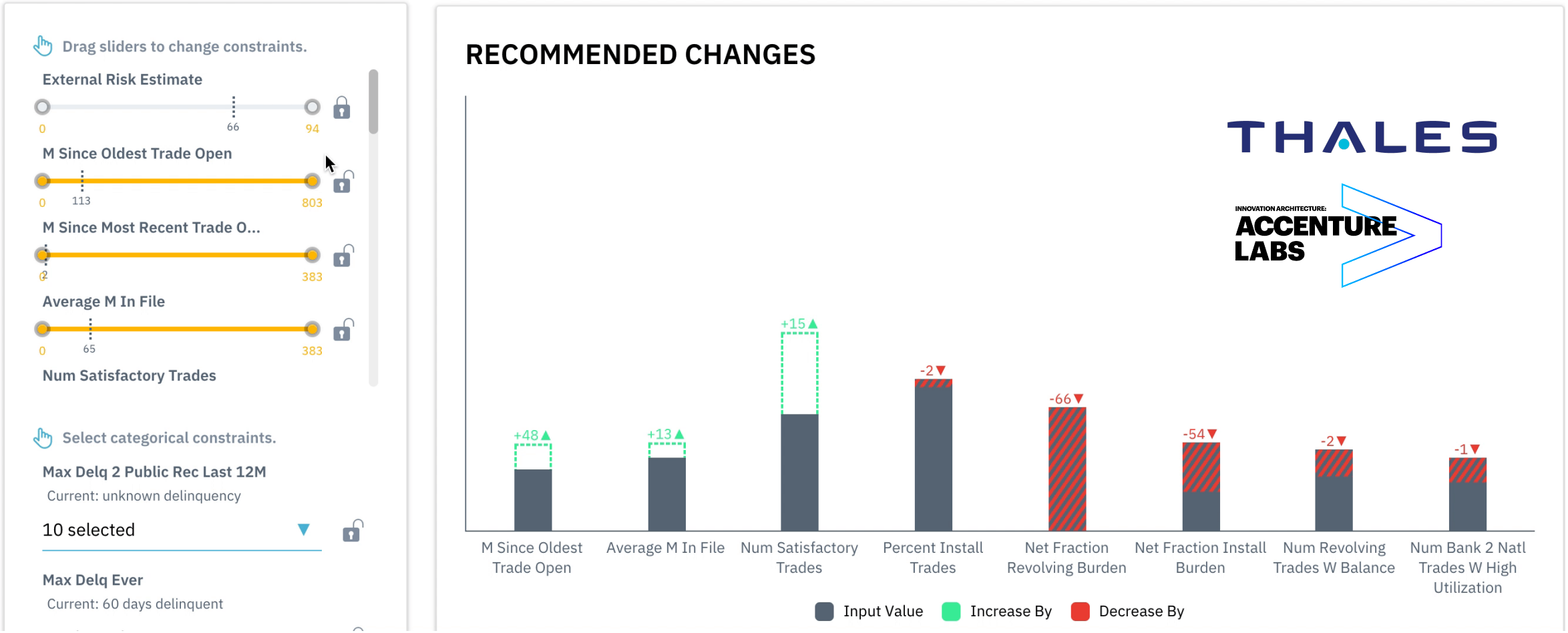
Rory Mc Grath, Luca Costabello, Chan Le Van, Paul Sweeney, Farbod Kamiab, Zhao Shen, Freddy Lécué: Interpretable Credit Application Predictions With Counterfactual Explanations. FEAP-AI4fin workshop, NeurIPS, 2018.

Counterfactual Explanations for Credit Decisions (2) - Finance



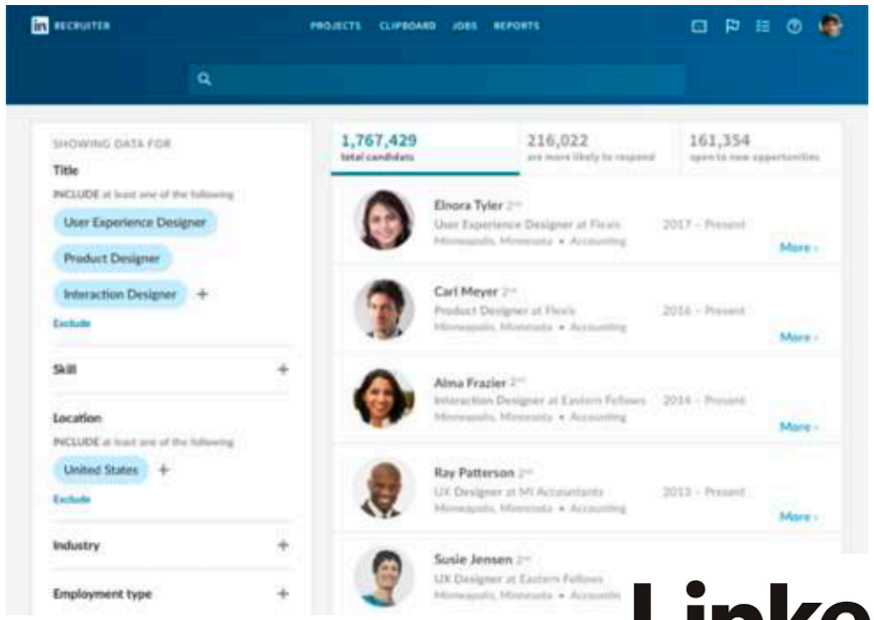
Rory Mc Grath, Luca Costabello, Chan Le Van, Paul Sweeney, Farbod Kamiab, Zhao Shen, Freddy Lécué: Interpretable Credit Application Predictions With Counterfactual Explanations. FEAP-AI4fin workshop, NeurIPS, 2018.

Counterfactual Explanations for Credit Decisions (3) - Finance

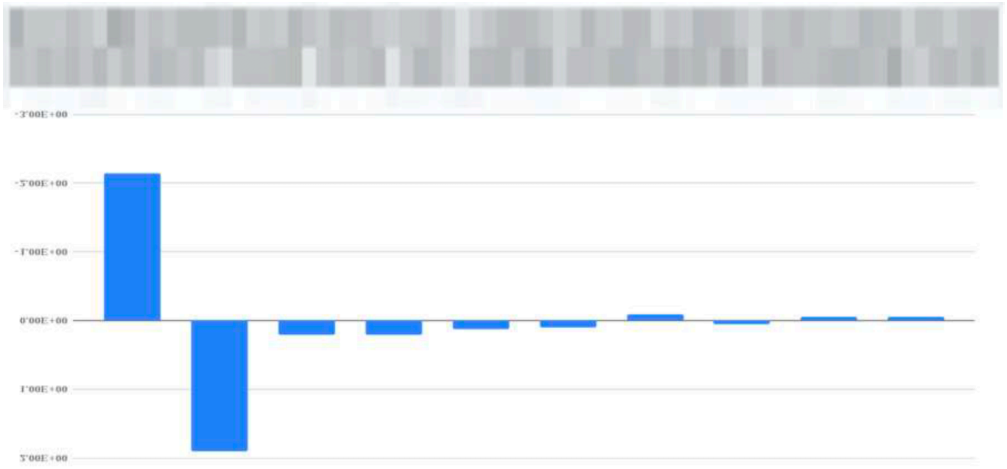


Rory Mc Grath, Luca Costabello, Chan Le Van, Paul Sweeney, Farbod Kamiab, Zhao Shen, Freddy Lécué: Interpretable Credit Application Predictions With Counterfactual Explanations. FEAP-AI4fin workshop, NeurIPS, 2018.

Explaining Talent Search Results – Human Resources



LinkedIn



Challenge: How to rationalize a talent search for a recruiter when looking for candidates for a given role. Features are dynamic and costly to compute. Recruiters are interested in discriminating between two candidates to make a selection.

AI Technology: Generalized Linear Mixed Models, Artificial Neural Networks, XGBoost

XAI Technology: Generalized Linear Mixed Models (inherently explainable), Integrated Gradient, Features Importance in XGBoost

Feature	Description	Difference (1 vs 2)	Contribution
Feature……	Description……	-2.0476928	-2.144455602
Feature……	Description……	-2.3223877	1.903594618
Feature……	Description……	0.11666667	0.2114946752
Feature……	Description……	-2.1442587	0.2060414469
Feature……	Description……	-14	0.1215354111
Feature……	Description……	1	0.1000282466
Feature……	Description……	-92	-0.085286277
Feature……	Description……	0.9333333	0.0568533262
Feature……	Description……	-1	-0.051796317
Feature……	Description……	-1	-0.050895940

Source: Explainable AI in Industry. KDD 2019 Tutorial. Varun Mithal, Girish Kathalagiri, Sahin Cem Geyik

Explanation of Medical Condition Relapse – Health

THALES

INNOVATION ARCHITECTURE: ACCENTURE LABS

Challenge: Explaining medical condition relapse in the context of oncology.

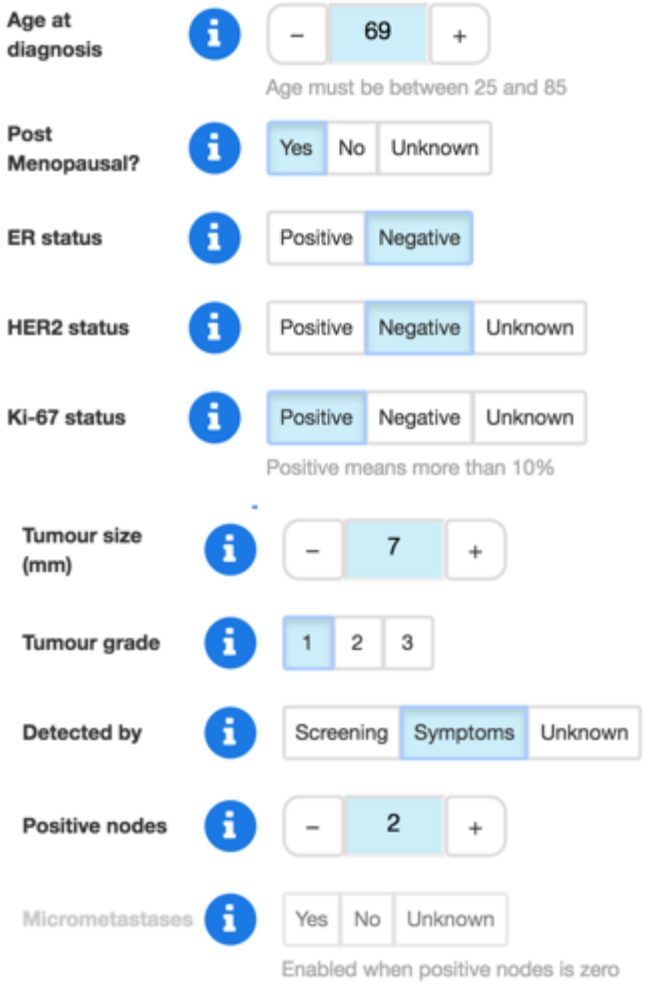
AI Technology: Relational learning

XAI Technology: Knowledge graphs and Artificial Neural Networks

Knowledge graph parts explaining medical condition relapse

Breast Cancer Survival Rate Prediction - Health

predict
breast cancer



Results

Table | Curves | Chart | Texts | Icons

New recording

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**

David Spiegelhalter, Making Algorithms trustworthy, NeurIPS 2018 Keynote
predict.nhs.uk/tool

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.

More on XAI

(Some) Tutorials, Workshops, Challenge

Tutorial:

- AAAI 2019 Tutorial on On Explainable AI: From Theory to Motivation, Applications and Limitations (#1) - https://xaitutorial2019.github.io/
- ICIP 2018 / EMBC 2019 Interpretable Deep Learning: Towards Understanding & Explaining Deep Neural Networks (#2) - http://interpretable-ml.org/icip2018tutorial/ - http://interpretable-ml.org/embc2019tutorial/
- ICCV 2019 Tutorial on Interpretable Machine Learning for Computer Vision (#2) - https://interpretablevision.github.io/
- KDD 2019 Tutorial on Explainable AI in Industry (#1) - https://sites.google.com/view/kdd19-explainable-ai-tutorial

Workshop:

- ISWC 2019 Workshop on Semantic Explainability (#1) - http://www.semantic-explainability.com/
- IJCAI 2019 Workshop on Explainable Artificial Intelligence (#3) - https://sites.google.com/view/xai2019/home 55 paper submitted in 2019
- IJCAI 2019 Workshop on Optimisation and Explanation in AI (#1) - https://www.doc.ic.ac.uk/~kc2813/OXAI/
- SIGIR 2019 Workshop on Explainable Recommendation and Search (#2) https://ears2019.github.io/
- ICAPS 2019 Workshop on Explainable Planning (#2)- https://kcl-planning.github.io/XAIP-Workshops/ICAPS_2019 23 papers submitted in 2019 https://openreview.net/group?id=icaps-conference.org/ICAPS/2019/Workshop/XAIP
- KDD 2019 Workshop on Explainable AI for fairness, accountability, and transparency (#1) – https://xai.kdd2019.a.intuit.com
- ICCV 2019 Workshop on Interpreting and Explaining Visual Artificial Intelligence Models (#1) - http://xai.unist.ac.kr/workshop/2019/
- NeurIPS 2019 Workshop on Challenges and Opportunities for AI in Financial Services: the Impact of Fairness, Explainability, Accuracy, and Privacy - https://sites.google.com/view/feap-ai4fin-2018/
- CD-MAKE 2019 – Workshop on Explainable AI (#2) - https://cd-make.net/special-sessions/make-explainable-ai/
- AAAI 2019 / CVPR 2019 Workshop on Network Interpretability for Deep Learning (#1 and #2) - http://networkinterpretability.org/ - https://explainai.net/
- IEEE FUZZ 2019 / Advances on eXplainable Artificial Intelligence (#2) - https://sites.google.com/view/xai-fuzzieee2019
- International Conference on NL Generation - Interactive Natural Language Technology for Explainable Artificial Intelligence (EU H2020 NL4XAI; #1) - https://sites.google.com/view/nl4xai2019/

Challenge:

- 2018: FICO Explainable Machine Learning Challenge (#1) - https://community.fico.com/s/explainable-machine-learning-challenge

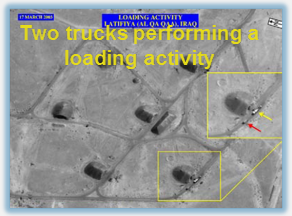
(Some) Software Resources

- DeepExplain: perturbation and gradient-based attribution methods for Deep Neural Networks interpretability. github.com/marcoancona/DeepExplain
- iNNvestigate: A toolbox to iNNvestigate neural networks' predictions. github.com/albermax/innvestigate
- SHAP: SHapley Additive exPlanations. github.com/slundberg/shap
- Microsoft Explainable Boosting Machines. https://github.com/Microsoft/interpret
- GANDissect: Pytorch-based tools for visualizing and understanding the neurons of a GAN. https://github.com/CSAILVision/GANDissect
- 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
- LIME: Agnostic Model Explainer. https://github.com/marcotcr/lime
- Sklearn_explain: model individual score explanation for an already trained scikit-learn model. https://github.com/antoinecarme/sklearn_explain
- Heatmapping: Prediction decomposition in terms of contributions of individual input variables
- Deep Learning Investigator: Investigation of Saliency, Deconvnet, GuidedBackprop and more. https://github.com/albermax/innvestigate
- Google PAIR What-if: Model comparison, counterfactual, individual similarity. https://pair-code.github.io/what-if-tool/
- Google tf-explain: https://tf-explain.readthedocs.io/en/latest/
- IBM AI Fairness: Set of fairness metrics for datasets and ML models, explanations for these metrics. https://github.com/IBM/aif360
- Blackbox auditing: Auditing Black-box Models for Indirect Influence. https://github.com/algofairness/BlackBoxAuditing
- Model describer: Basic statiscal metrics for explanation (visualisation for error, sensitivity). https://github.com/DataScienceSquad/model-describer
- *AXA Interpretability and Robustness: https://axa-rev-research.github.io/ (more on research resources – not much about tools)*

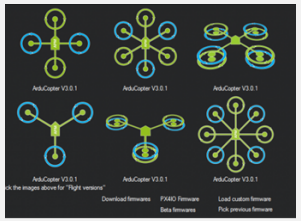
(Some) Initiatives: XAI in USA

DARPA

Challenge Problem Areas



Data Analytics
Multimedia Data

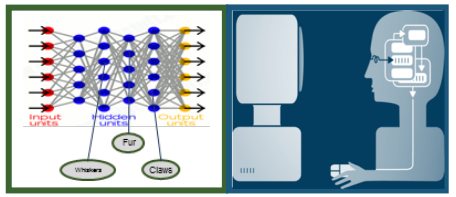


Autonomy
ArduPilot &
SITL Simulation

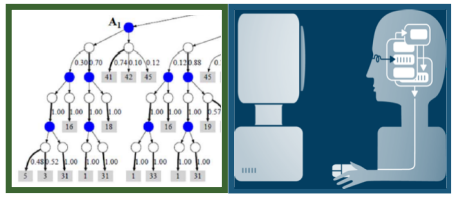
TA 1: Explainable Learners

Teams that provide prototype systems with both components:

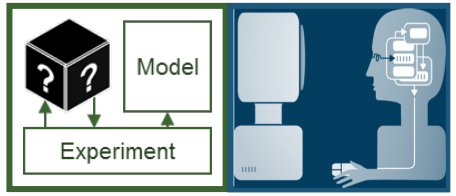
- Explainable Model
- Explanation Interface



Deep Learning Teams



Interpretable Model Teams

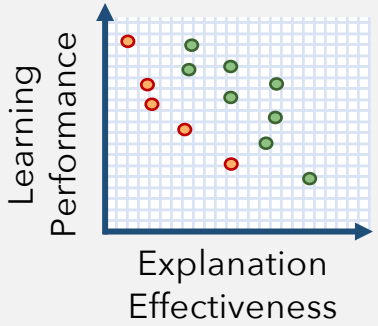


Model Induction Teams

TA 2: Psychological Model of Explanation

- Psych. Theory of Explanation
- Computational Model
- Consulting

Evaluation Framework



Explanation Measures

- User Satisfaction
- Mental Model
- Task Performance
- Trust Assessment
- Correctability

Evaluator

TA1: Explainable Learners

- Explainable learning systems that include both an explainable model and an explanation interface

TA2: Psychological Model of Explanation

- Psychological theories of explanation and develop a computational model of explanation from those theories

(Some) Initiatives: XAI in Canada

- DEEL CRIAQ ab IVADO in CRSNG NSERC arning) Project 2019-2024
  - Research institutions
  - Industrial partners
  - A ers

Bell Helicopter A Textron Company · BOMBARDIER · CAE · THALES

IRT SAINT EXUPERY · nnolo POLYTECHNIQUE MONTRÉAL · w met UNIVERSITÉ LAVAL s Trustable and Explainable AI

System Robustness
- To biased data
- Of algorithm
- To change
- To attacks

Certificability
- Structural warranties
- Risk auto evaluation
- External audit

Explicability & Interpretability

Privacy by design
- Differential privacy
- Homomorphic coding
- Collaborative learning
- To attacks

(Some) Initiatives: XAI in EU

Conclusion

Why do we Need XAI by the Way?

- ***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
- **To scale and industrialize** AI

Conclusion

- Explainable AI is motivated by **real-world applications in AI**
- Not a new problem – a reformulation of past research challenges in AI
- Multi-disciplinary: multiple AI fields, HCI, social sciences (multiple definitions)
- In AI (in general): many interesting / complementary approaches
- **Many industrial applications already – crucial for AI adoption in critical systems**

Open Research Questions

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

Future Challenges

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

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

Job Openings

Research and Technology Applied AI (Artificial Intelligence) Scientist

Wherever safety and Security are Critical, Thales c
build smarter solutions. Everywhere.

s is a global technology leader for the Defen
l-class technology, the combined expertise c
have made Thales a key player in keeping the pub
protecting the national security interests of count

Established in 1972, Thales Canada has over 1,80
Toronto and Vancouver working in Defence, Avior

This is a unique opportunity to play a key role on t
Technology (TRT) in Canada (Quebec and Montrea
applied R&T experts at five locations worldwide. T
intelligence technologies. Our passion is imagining
cutting edge AI technologies. Not only will you joi
network, but this TRT is also co-located within Cor
Intelligence eXpertise) i.e., the new flagship progr
to work.

Job Description

An AI (Artificial Intelligence) Research and Techno
developing innovative prototypes to demonstrate
intelligence. To be successful in this role, one mos
what's new, and a strong ability to learn new tech
hand-on technical skills and be familiar with latest
will contribute as technical subject matter experts
and its business units. In addition to the impleme
individual will also be involved in the initial projec
thinking, and team work is also critical for this role

As a Research and Technology Applied AI Scientist
paced projects.

Professional Skill Requirements

- Good foundation in mathematics, statistic
- Strong knowledge of Machine Learning foundations
- Strong development skills with Machine Learning frameworks e.g., Scikit-learn, Tensoflow, PyTorch, Theano
- Knowledge of mainstream Deep Learning architectures (MLP, CNN, RNN, etc).
- Strong Python programming skills
- Working knowledge of Linux OS
- Eagerness to contribute in a team-oriented environment
- Demonstrated leadership abilities in school, civil or business organisations
- Ability to work creatively and analytically in a problem-solving environment
- Proven verbal and written communication skills in English (talks, presentations, publications, etc.)

Basic Qualifications

- Master's degree in computer science, engineering or mathematics fields
- Prior experience in artificial intelligence, machine learning, natural language processing, or advanced analytics

Preferred Qualifications

- Minimum 3 years of analytic experience Python with interest in artificial intelligence with working structured and unstructured data (SQL, Cassandra, MongoDB, Hive, etc.)
- A track record of outstanding AI software development with Github (or similar) evidence
- Demonstrated abilities in designing large scale AI systems
- Demonstrated interes in Explainable AI and or relational learning
- Work experience with programming languages such as C, C++, Java, scripting languages (Perl/Python/Ruby) or similar
- Hands-on experience with data visualization, analytics tools/languages
- Demonstrated teamwork and collaboration in professional settings
- Ability to establish credibility with clients and other team members

AUGUST 28TH, 2019

Freddy Lecue
Chief AI Scientist, CortAIx, Thales, Montreal – Canada

@freddylecue
https://tinyurl.com/freddylecue
Freddy.lecue.e@thalesdigital.io