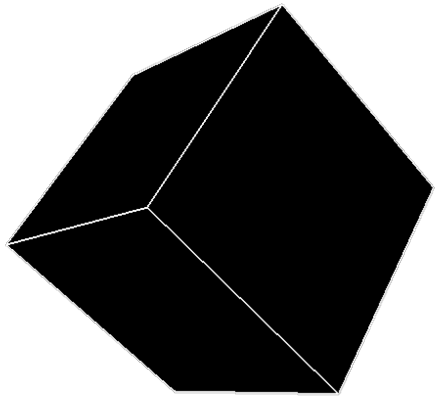
XAI - Explanation in AI: From Machine Learning to Knowledge Representation & Reasoning and Beyond

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

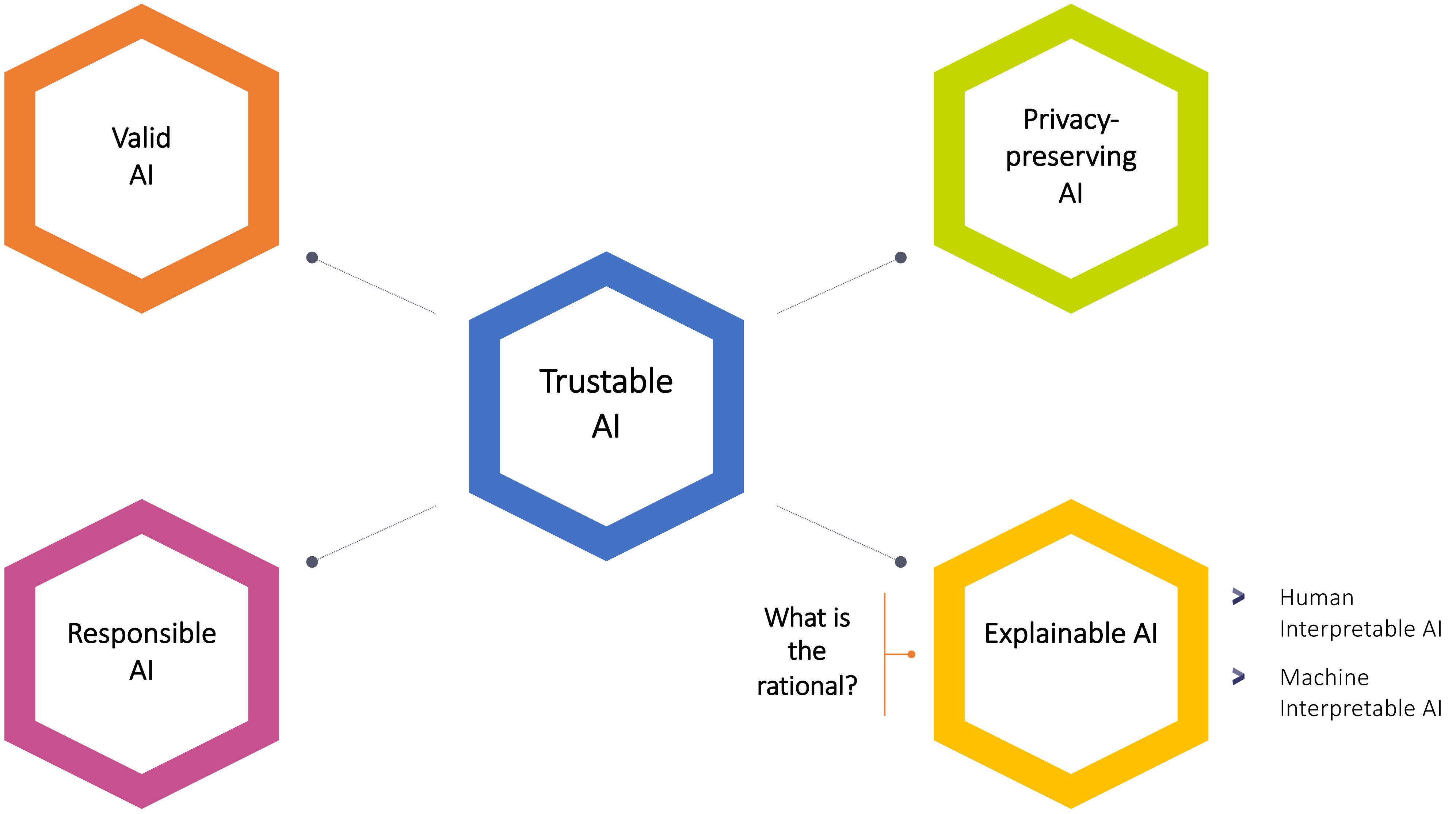
March 4th, 2020
AI@Centech
Centech, Montreal, Quebec, Canada



Freddy Lecue. On the Role of Knowledge Graphs in Explainable AI. Semantic Web Journal (to appear 2020) http://www.semantic-web-journal.net/content/role-knowledge-graphs-explainable-ai

Scope

AI Adoption: Requirements



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

Outline

- **Explanation in Artificial Intelligence**
  - Motivation
  - Definitions
  - Evaluation (with role of the human in XAI systems)
  - The Role of Humans
  - Explanations in Different AI fields
- **On the Role of Knowledge Graph in Explainable Machine Learning**
- **XAI Industrial Applications using Knowledge Graphs on Machine Learning**
- **Conclusion + Q&A**

Motivation

Business to Customer

Critical Systems


BNSF
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

Motivation (2)

Finance:

- Credit scoring, loan approval
- Insurance quotes

FICO COMMUNITY

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 3^{rd}-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

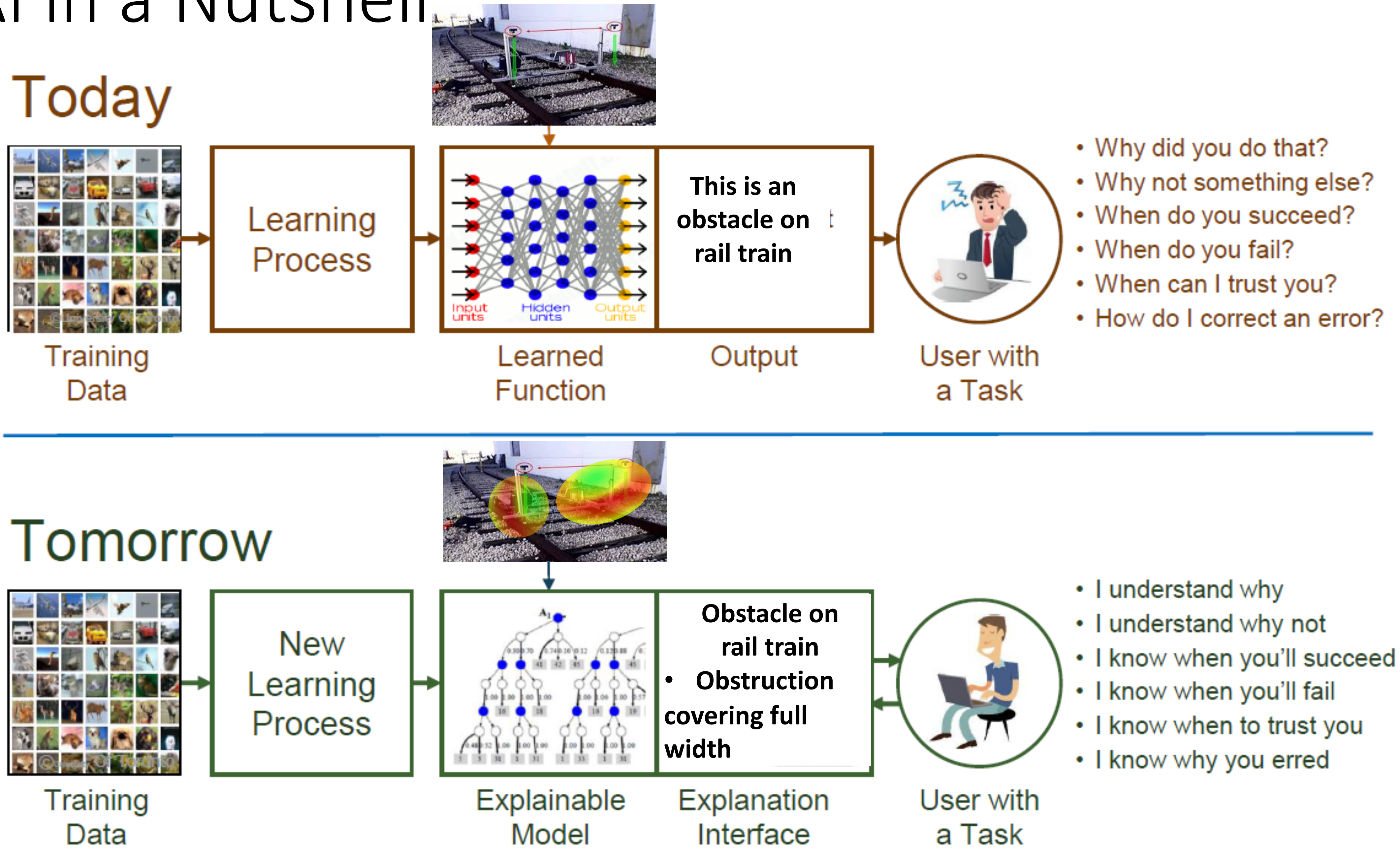
Marc Sturm
NewYork-Presbyterian Hospital
mas9161@nyp.org

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

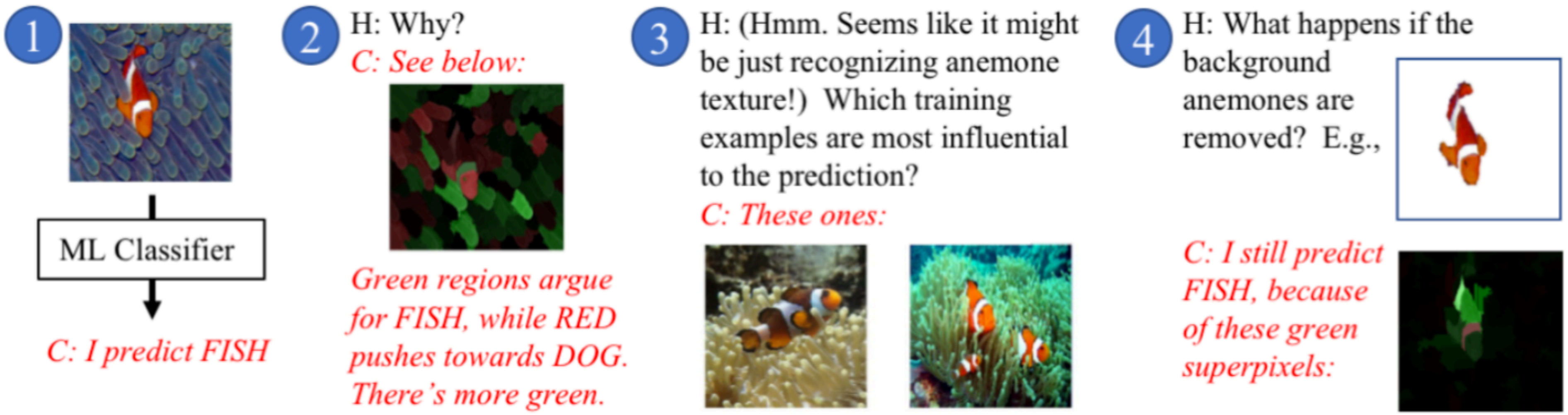
XAI in a Nutshell

XAI in a Nutshell



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

An Example of an end-to-end XAI System



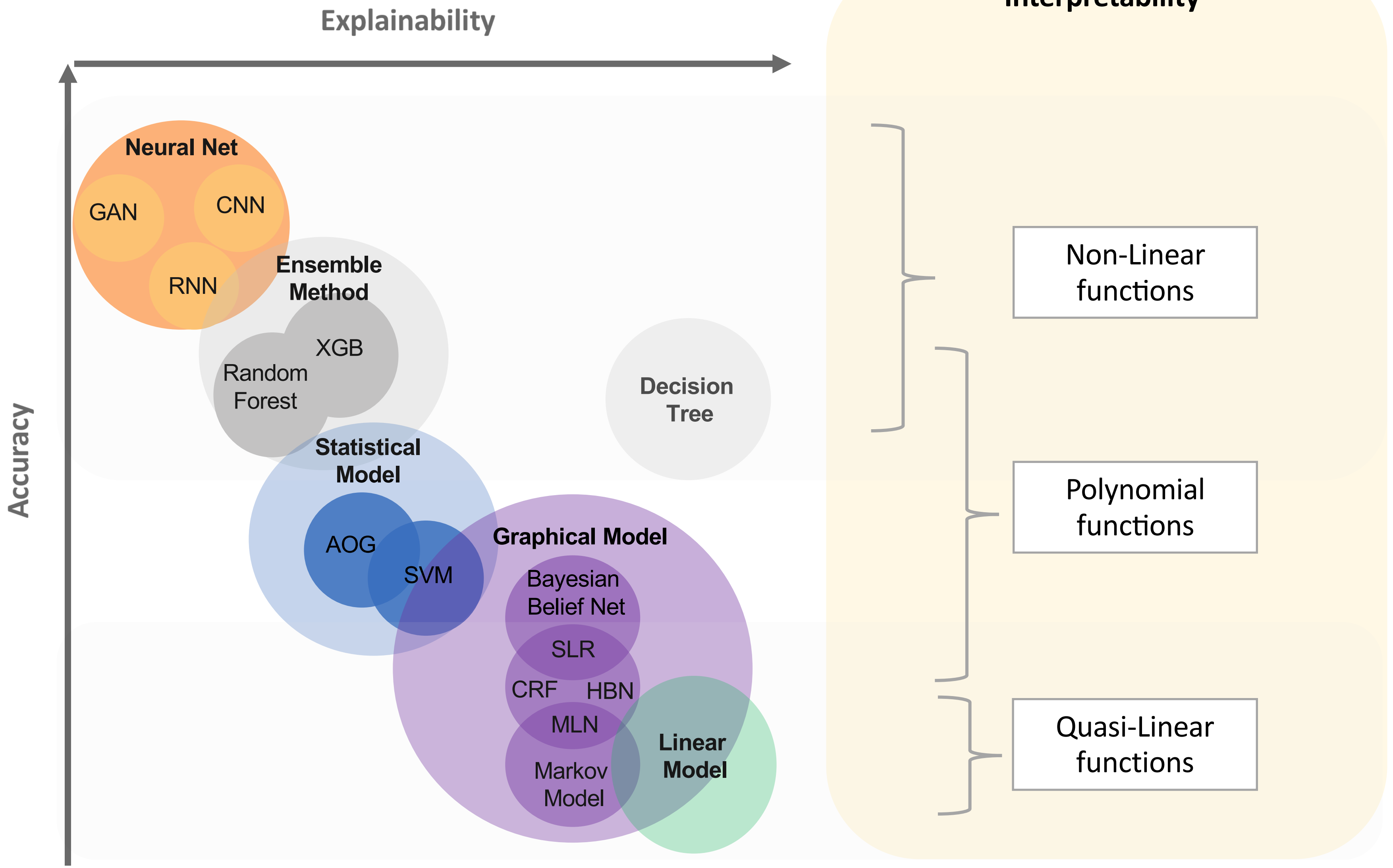
- Humans may have follow-up questions
- Explanations cannot answer all users' concerns

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

How to Explain? Accuracy vs. Explanability

Learning

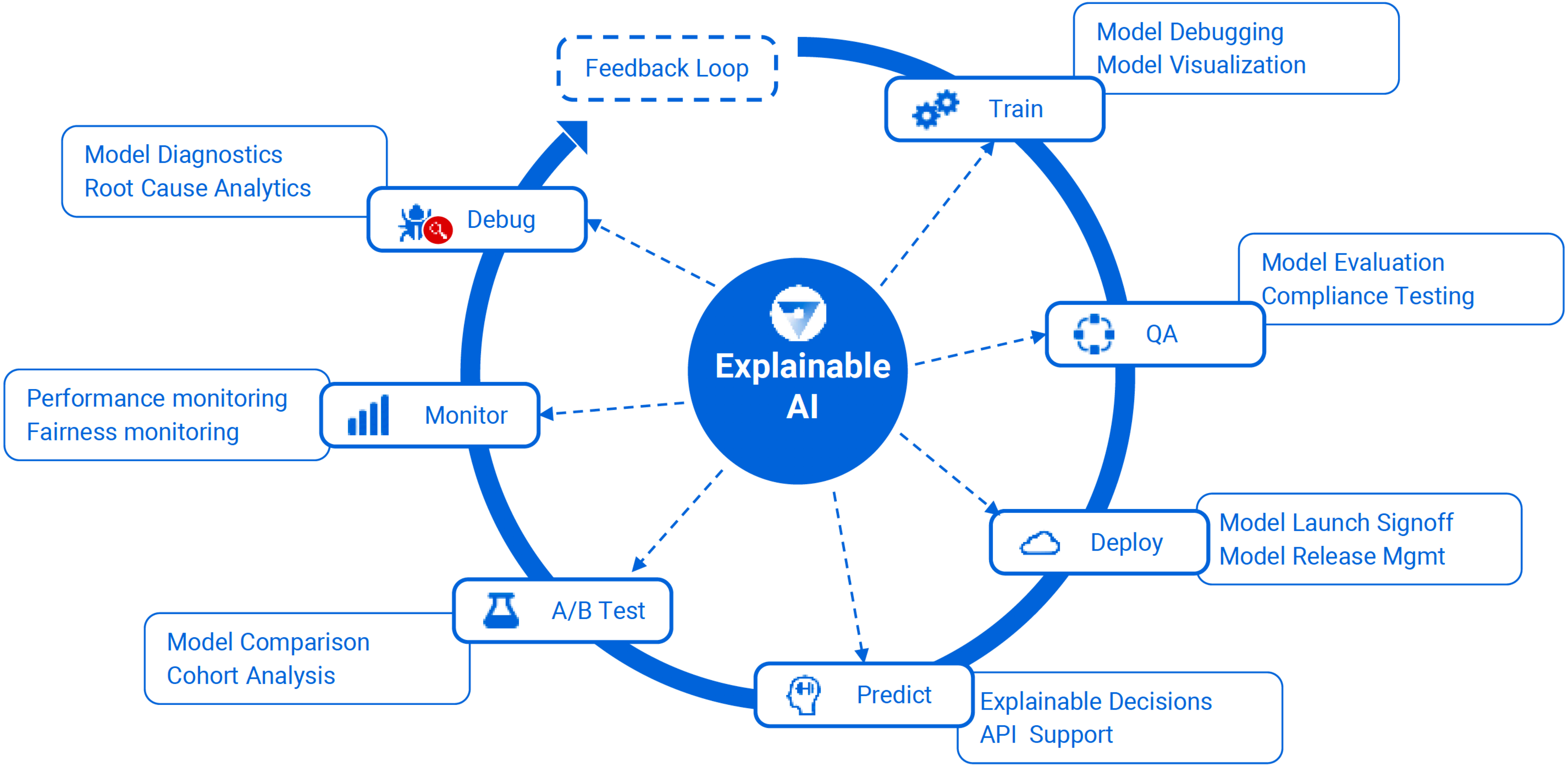
- Challenges:
  - Supervised
  - Unsupervised learning
- Approach:
  - Representation Learning
  - Stochastic selection
- Output:
  - **Correlation**
  - **No causation**



XAI Objective

Supporting Industrialization of AI at Scale

Explainability by Design for AI Products



XAI Definitions

Explanation vs. Interpretability

Oxford Dictionary of English

explanation | ɛkspləˈneɪʃ(ə)n |

noun

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

interpret | ɪnˈtəːprɪt |

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

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

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

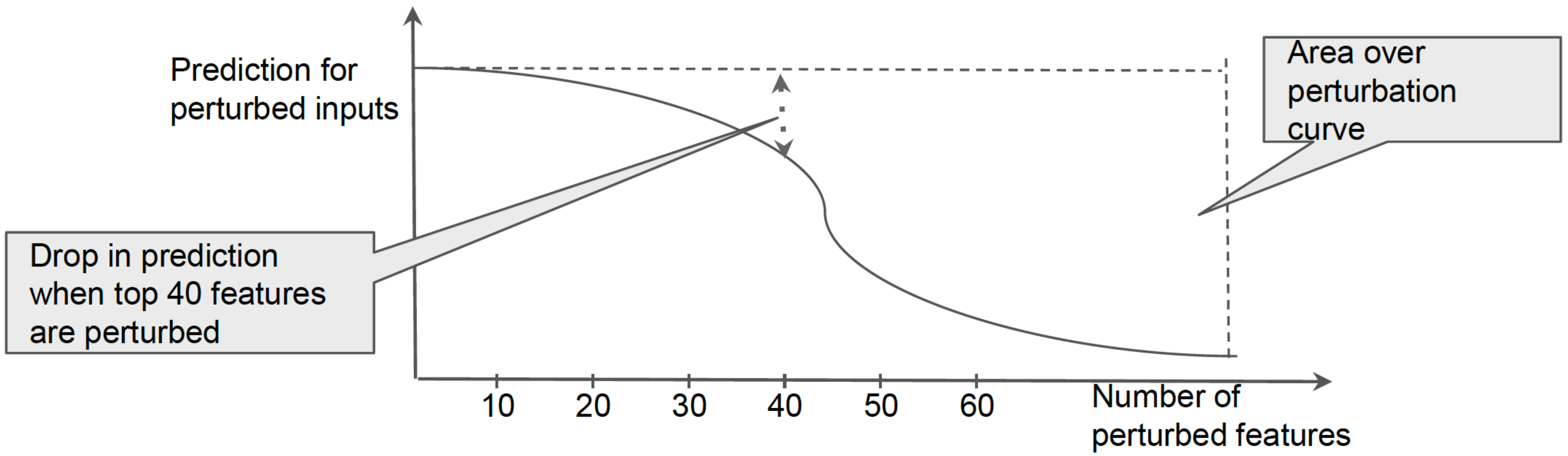
Text

What about the Evaluation?

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



Human (Role)-based Evaluation is Essential… but too often based on size!

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]

XAI: One Objective, Many Metrics

Comprehensibility

How much effort for correct human interpretation?

Succinctness

How concise and compact is the explanation?

Actionability

What can one action, do with the explanation?

Reusability

Could the explanation be personalized?

Accuracy

How accurate and precise is the explanation?

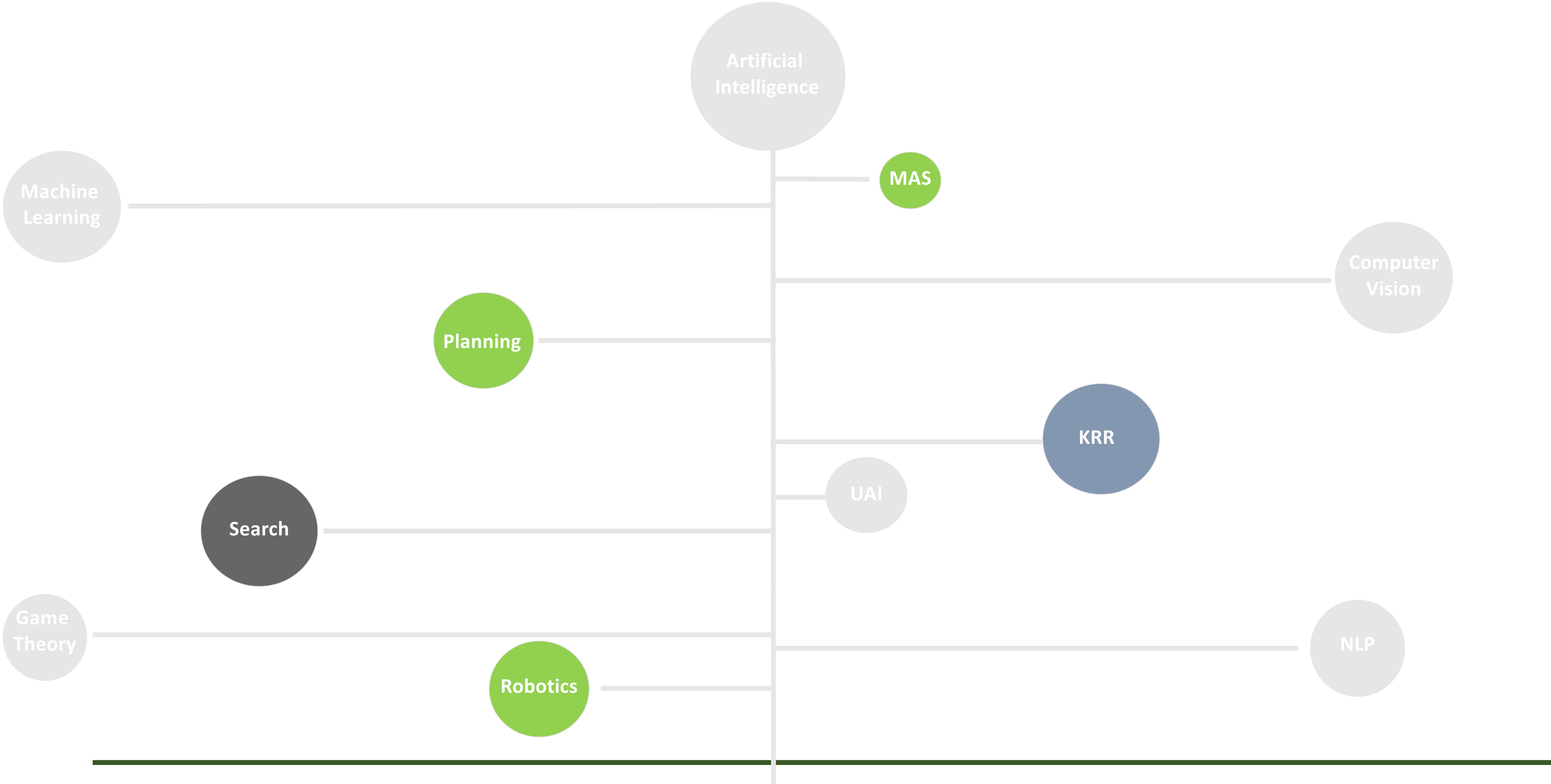
Completeness

Is the explanation complete, partial, restricted?

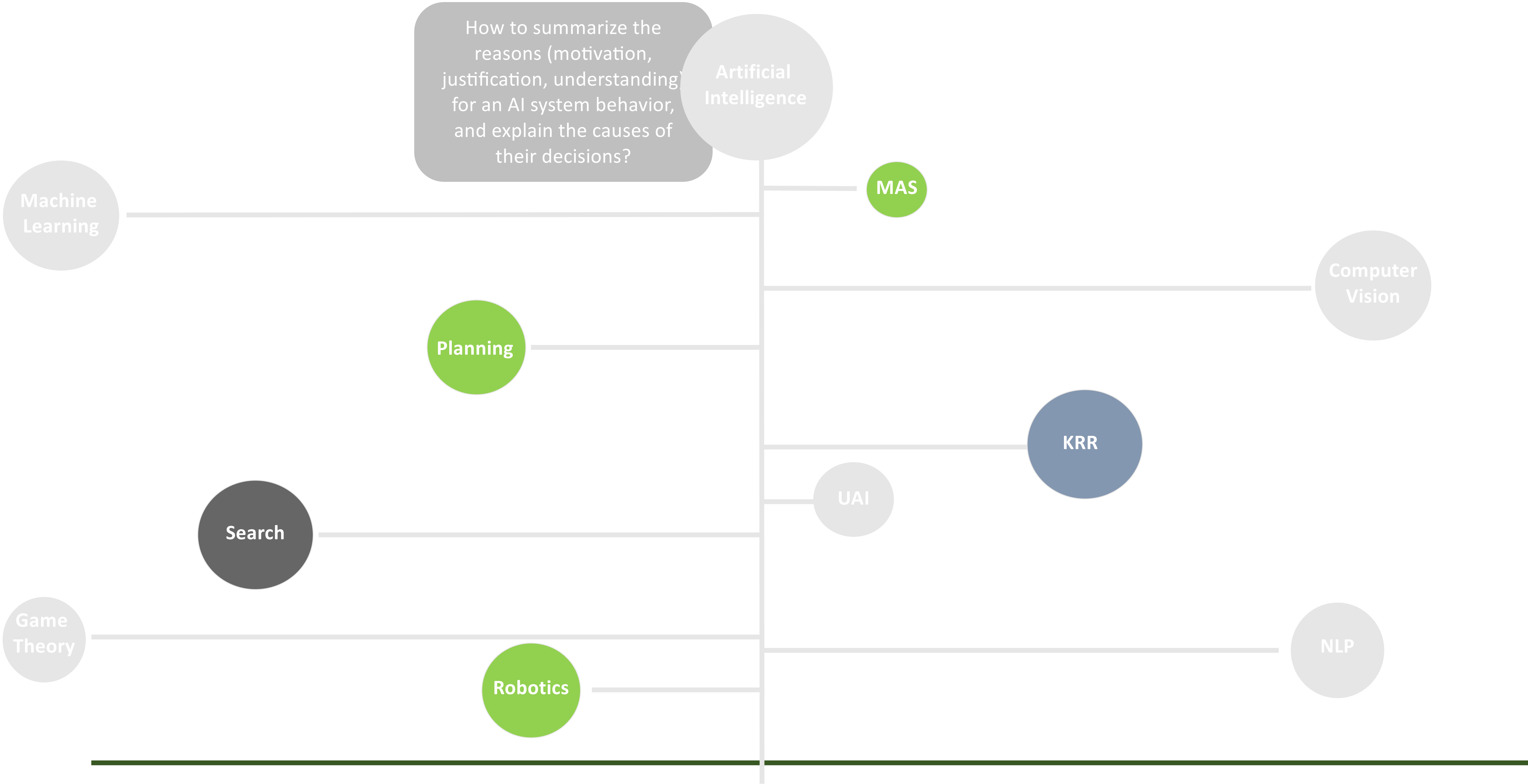
Source: Accenture Point of View. Understanding Machines: Explainable AI. Freddy Lecue, Dadong Wan

XAI in AI

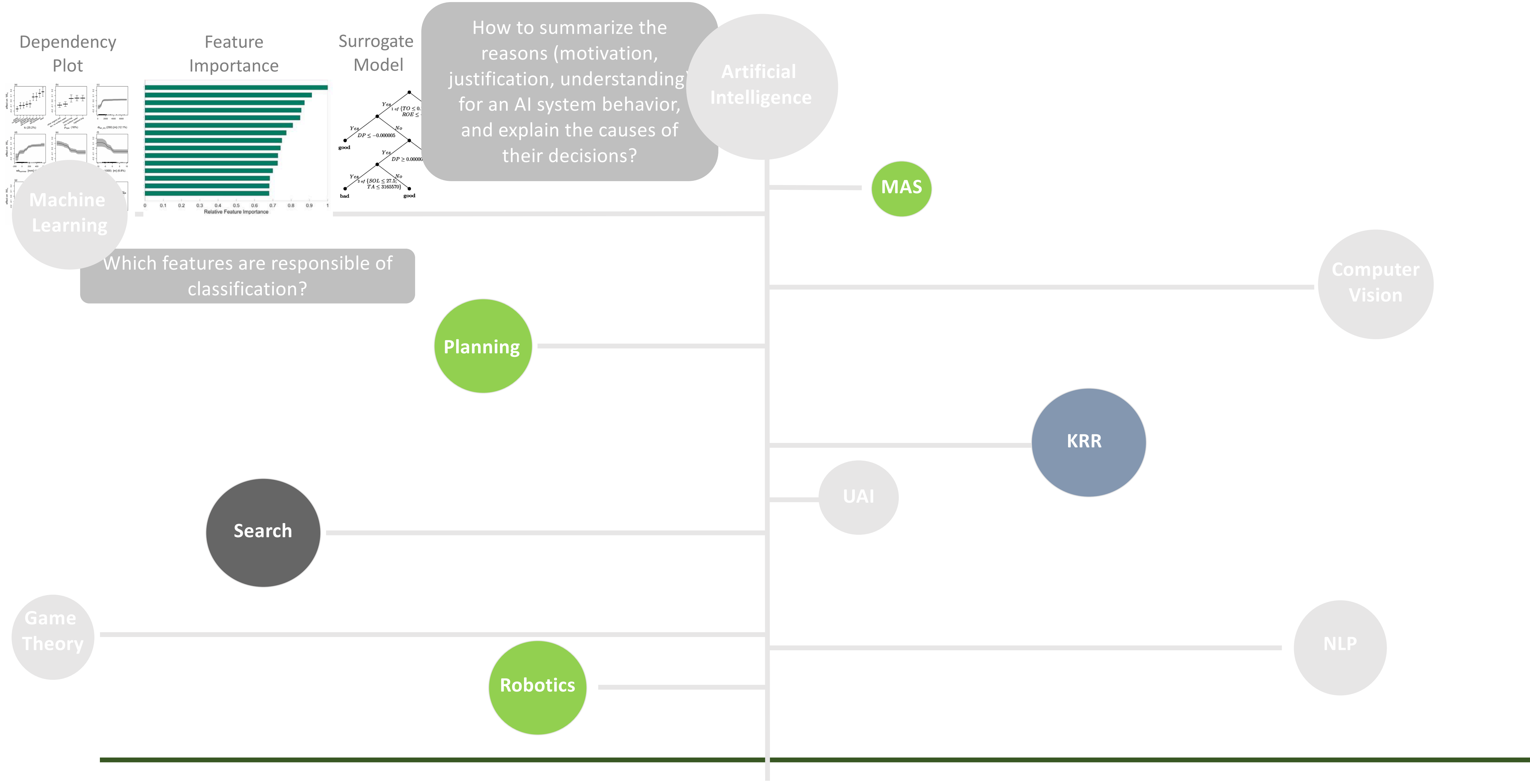
XAI: One Objective, Many 'AI's, Many Definitions, Many Approaches



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



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

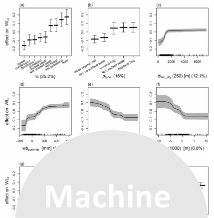


XAI: One Objective, Many 'AI's, Many Definitions, Many Approaches

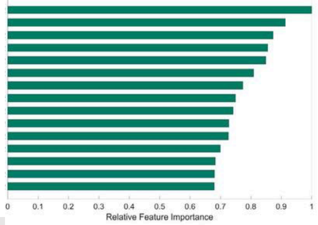
How to summarize the reasons (motivation, justification, understanding) for an AI system behavior, and explain the causes of their decisions?

Artificial Intelligence

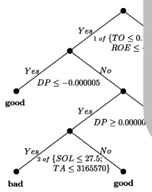
Dependency Plot



Feature Importance



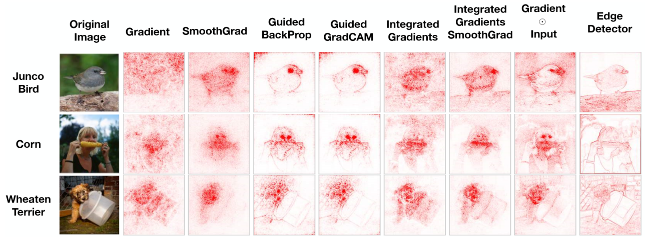
Surrogate Model



Machine Learning

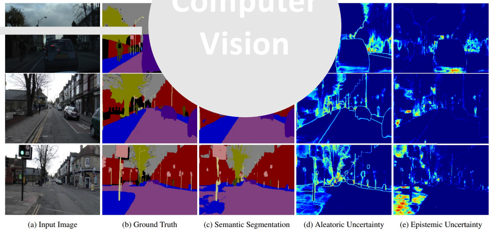
Which features are responsible of classification?

Saliency Map



Which complex features are responsible of classification?

Computer Vision



Uncertainty Map

MAS

Planning

KRR

UAI

Search

Game Theory

NLP

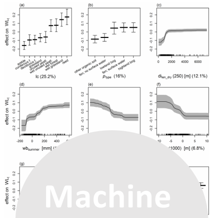
Robotics

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

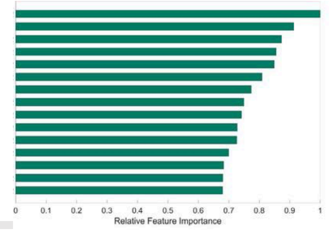
How to summarize the reasons (motivation, justification, understanding) for an AI system behavior, and explain the causes of their decisions?

Artificial Intelligence

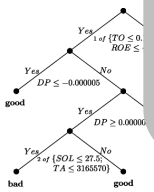
Dependency Plot



Feature Importance



Surrogate Model



Machine Learning

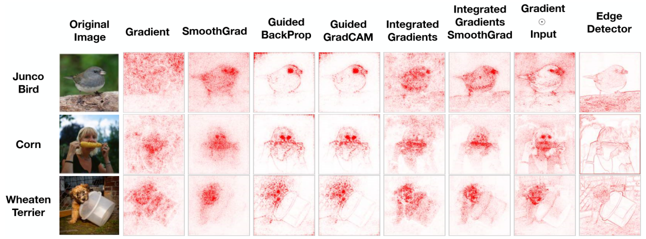
Which features are responsible of classification?

Strategy Summarization

MAS

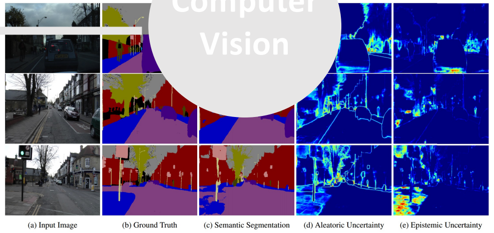
- Which agent strategy & plan ?
- Which player contributes most?
- Why such a conversational flow?

Saliency Map



Which complex features are responsible of classification?

Computer Vision



Uncertainty Map

Planning

KRR

UAI

Search

Game Theory

NLP

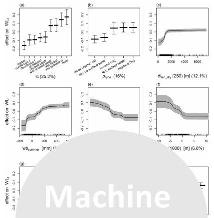
Robotics

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

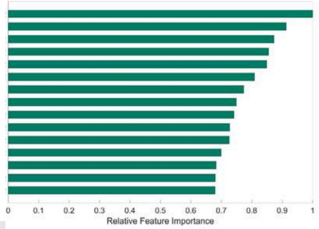
How to summarize the reasons (motivation, justification, understanding) for an AI system behavior, and explain the causes of their decisions?

Artificial Intelligence

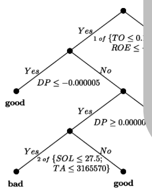
Dependency Plot



Feature Importance



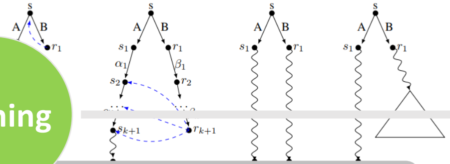
Surrogate Model



Machine Learning

Which features are responsible of classification?

Plan Refinement



Planning

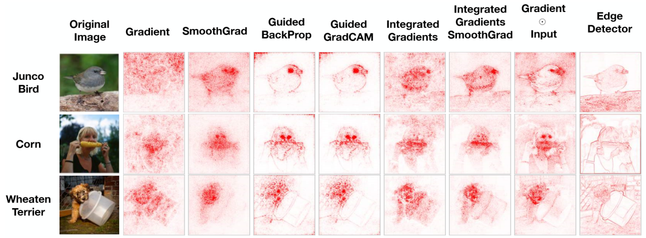
Which actions are responsible of a plan?

Strategy Summarization

MAS

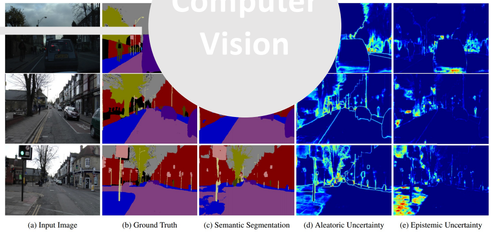
- Which agent strategy & plan ?
- Which player contributes most?
- Why such a conversational flow?

Saliency Map



Which complex features are responsible of classification?

Computer Vision



Uncertainty Map

KRR

UAI

Search

Game Theory

Robotics

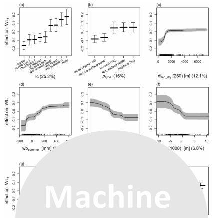
NLP

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

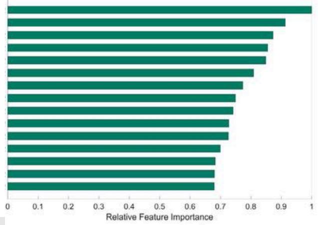
How to summarize the reasons (motivation, justification, understanding) for an AI system behavior, and explain the causes of their decisions?

Artificial Intelligence

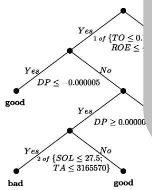
Dependency Plot



Feature Importance



Surrogate Model



Machine Learning

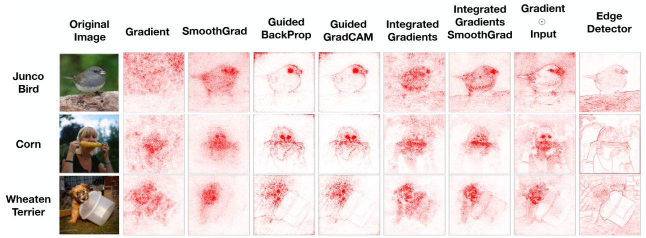
Which features are responsible of classification?

Strategy Summarization

MAS

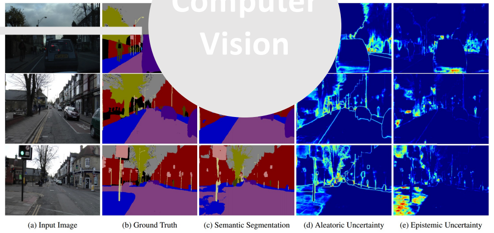
- Which agent strategy & plan ?
- Which player contributes most?
- Why such a conversational flow?

Saliency Map



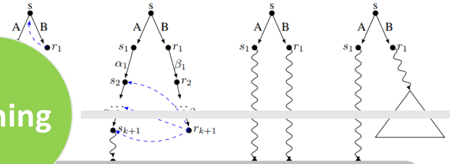
Which complex features are responsible of classification?

Computer Vision



Uncertainty Map

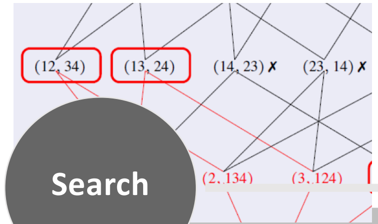
Plan Refinement



Planning

Which actions are responsible of a plan?

Conflicts Resolution



Search

Which constraints can be relaxed?

KRR

UAI

Game Theory

Robotics

NLP

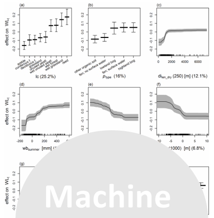
XAI: One Objective, Many 'AI's, Many Definitions, Many Approaches

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

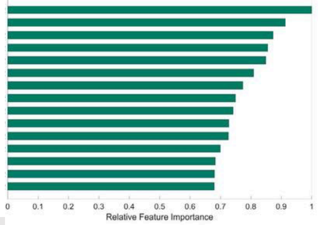
Artificial Intelligence

Machine Learning

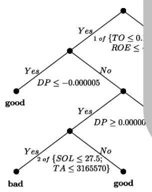
Dependency Plot



Feature Importance



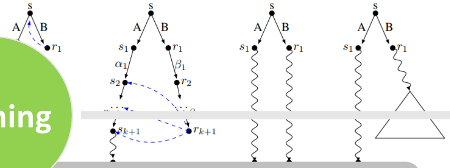
Surrogate Model



Which features are responsible of classification?

Planning

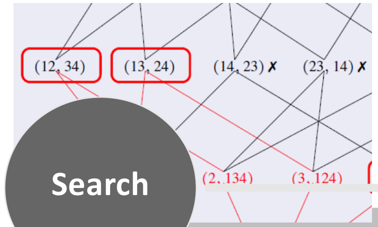
Plan Refinement



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

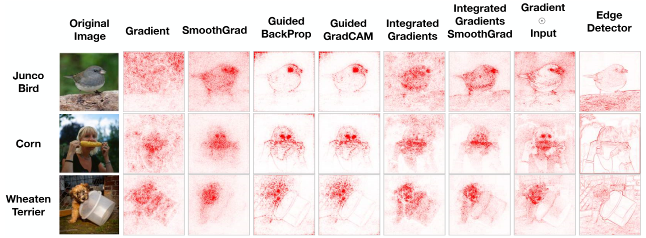
Robotics

MAS

Strategy Summarization

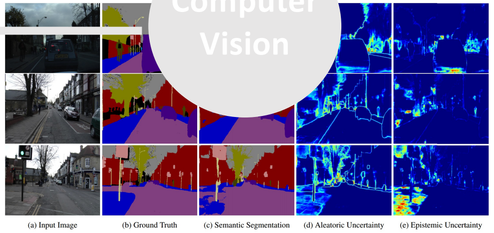
- Which agent strategy & plan ?
- Which player contributes most?
- Why such a conversational flow?

Saliency Map



Which complex features are responsible of classification?

Computer Vision



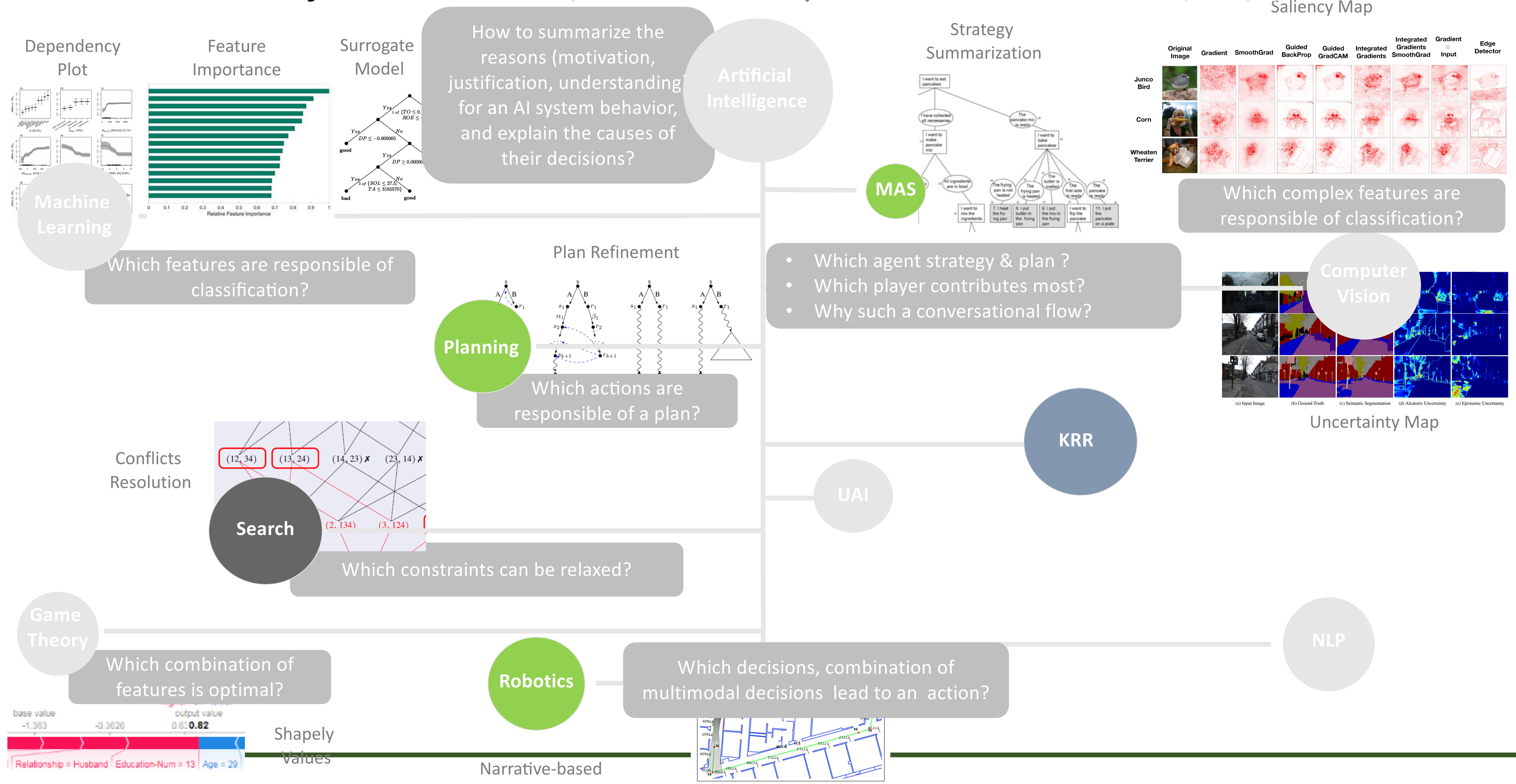
Uncertainty Map

KRR

UAI

NLP

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

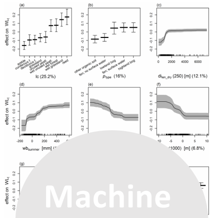


XAI: One Objective, Many 'AI's, Many Definitions, Many Approaches

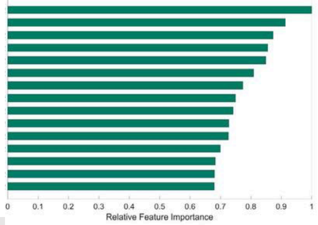
How to summarize the reasons (motivation, justification, understanding) for an AI system behavior, and explain the causes of their decisions?

Artificial Intelligence

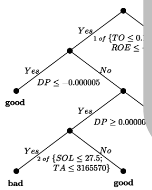
Dependency Plot



Feature Importance



Surrogate Model



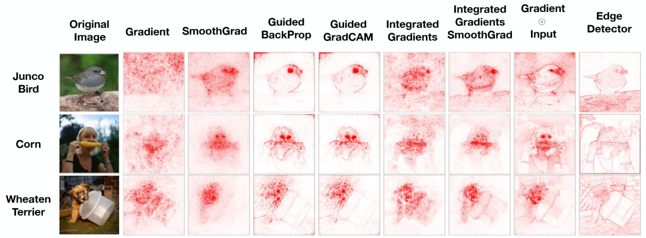
Machine Learning

Which features are responsible of classification?

Strategy Summarization

MAS

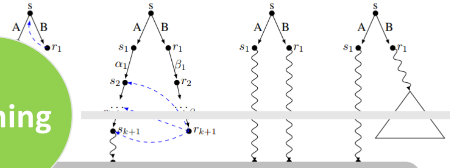
Saliency Map



Which complex features are responsible of classification?

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

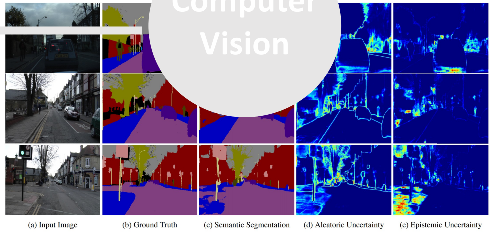
Plan Refinement



Planning

Which actions are responsible of a plan?

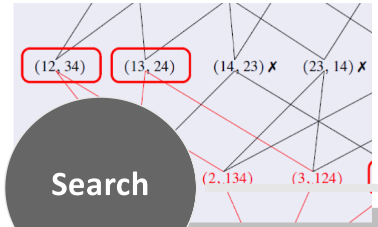
Computer Vision



Uncertainty Map

KRR

Conflicts Resolution



Search

Which constraints can be relaxed?

UAI

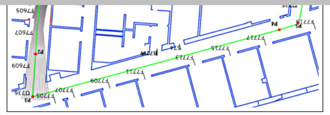
Game Theory

Which combination of features is optimal?

Shapely Values

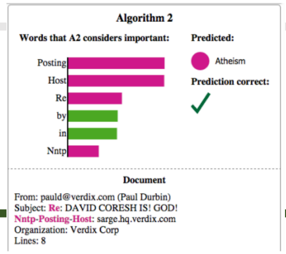
Robotics

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



Narrative-based

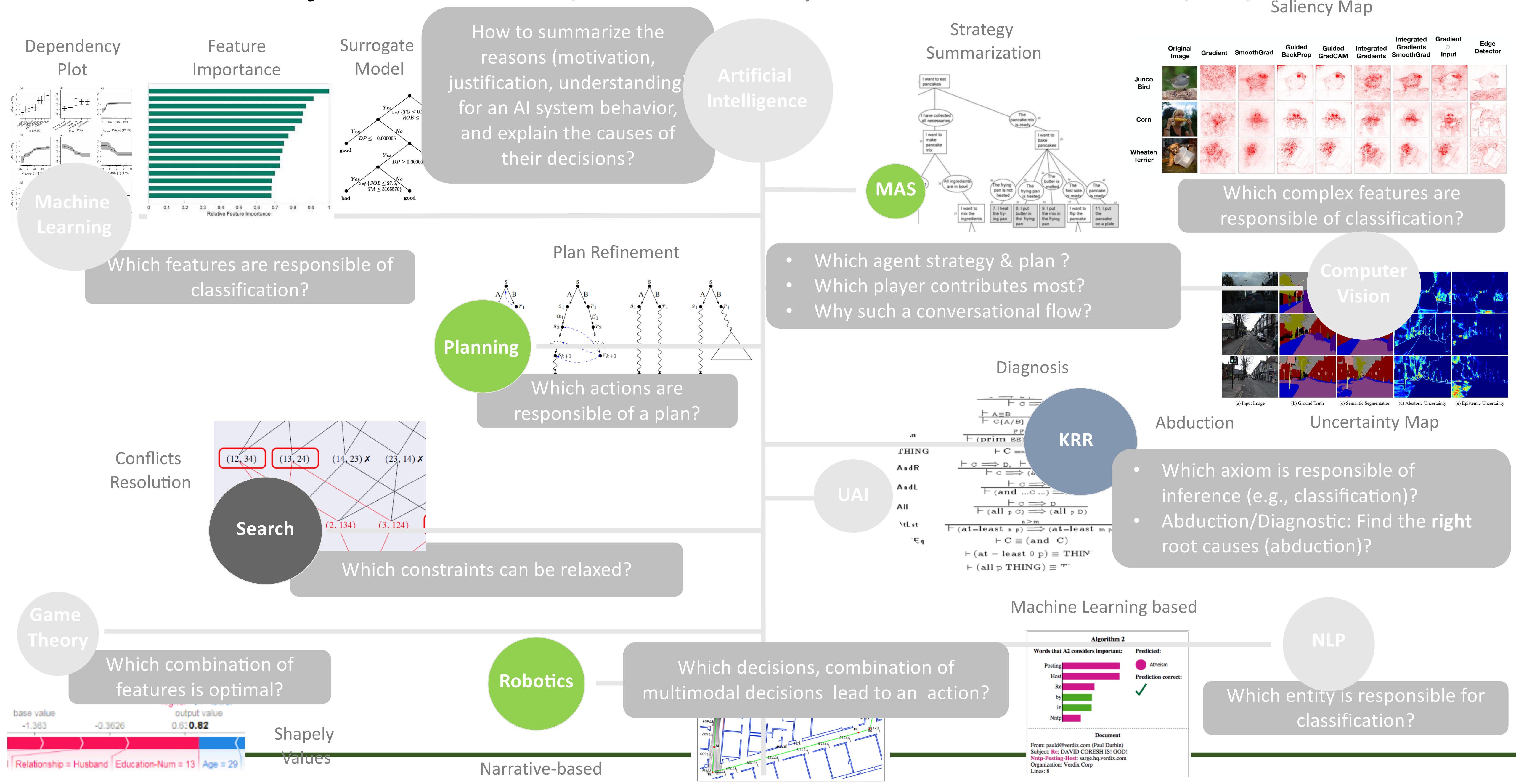
Machine Learning based



NLP

Which entity is responsible for classification?

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

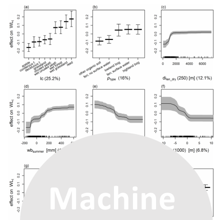


XAI: One Objective, Many 'AI's, Many Definitions, Many Approaches

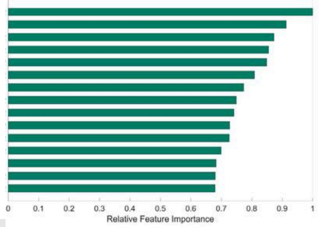
Artificial Intelligence

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

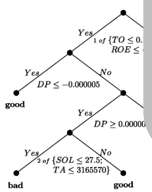
Dependency Plot



Feature Importance



Surrogate Model



Machine Learning

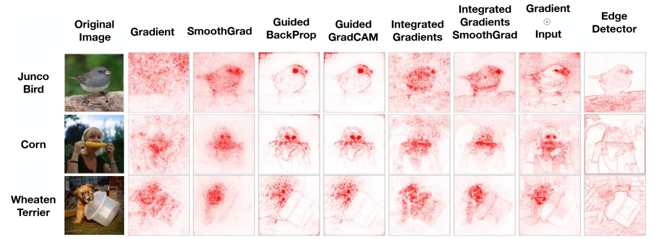
Which features are responsible of classification?

Strategy Summarization

MAS

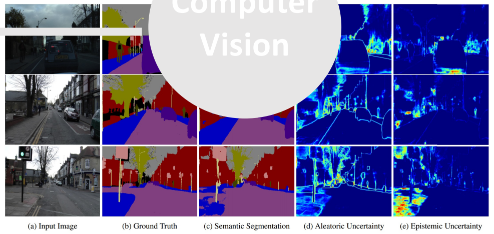
- Which agent strategy & plan ?
- Which player contributes most?
- Why such a conversational flow?

Saliency Map



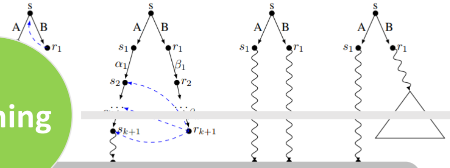
Which complex features are responsible of classification?

Computer Vision



Uncertainty Map

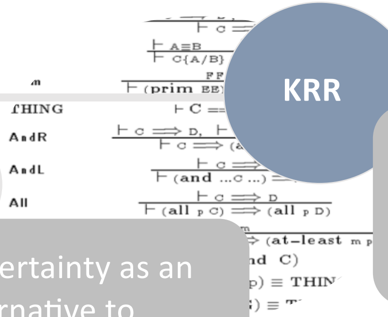
Plan Refinement



Planning

Which actions are responsible of a plan?

Diagnosis

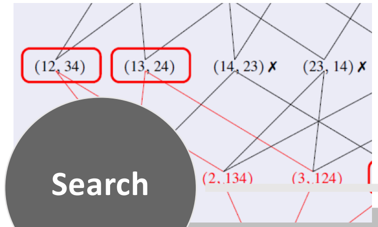


KRR

Abduction

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

Conflicts Resolution



Search

Which constraints can be relaxed?

UAI

Uncertainty as an alternative to explanation

Machine Learning based

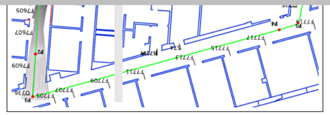
Game Theory

Which combination of features is optimal?

Shapely Values

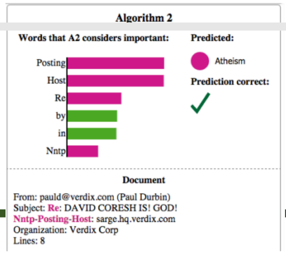
Robotics

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



Narrative-based

NLP



Which entity is responsible for classification?

Deep Dive

Overview of explanation in different AI fields (1)

- Machine Learning (except Artificial Neural Network)

Interpretable Models:

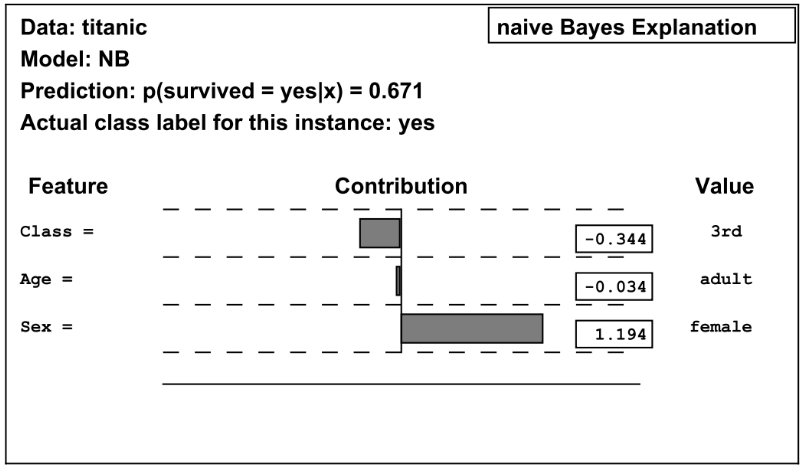
- Linear regression,
- Logistic regression,
- Decision Tree,
- GLMs,
- GAMs
- KNNs

Overview of explanation in different AI fields (1)

- Machine Learning (except Artificial Neural Network)

Interpretable Models:

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



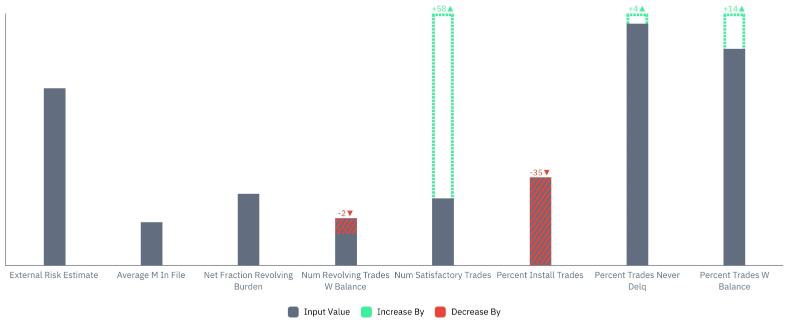
Naive Bayes model

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 different AI fields (1)

- Machine Learning (except Artificial Neural Network)

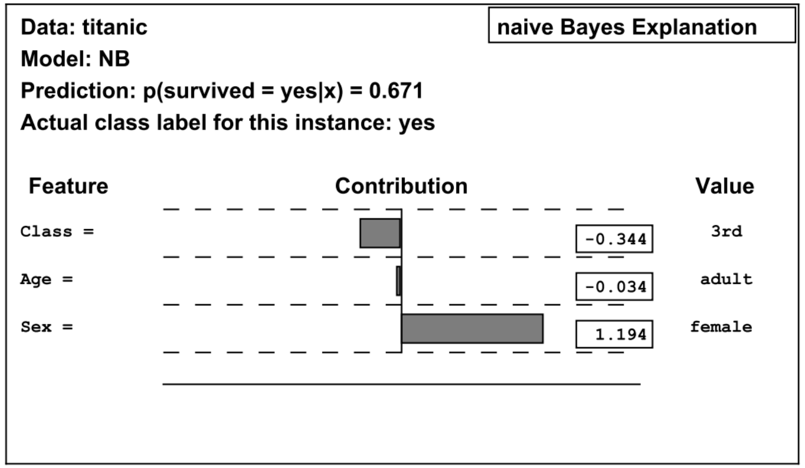
Interpretable Models:
- Linear regression,
- Logistic regression,
- Decision Tree,
- GLMs,
- GAMs
- KNNs



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)



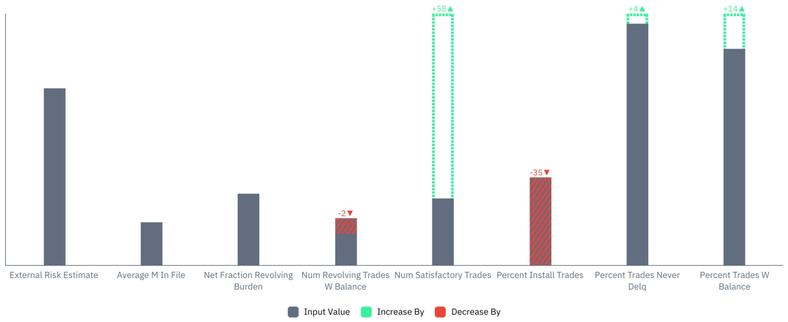
Naive Bayes model

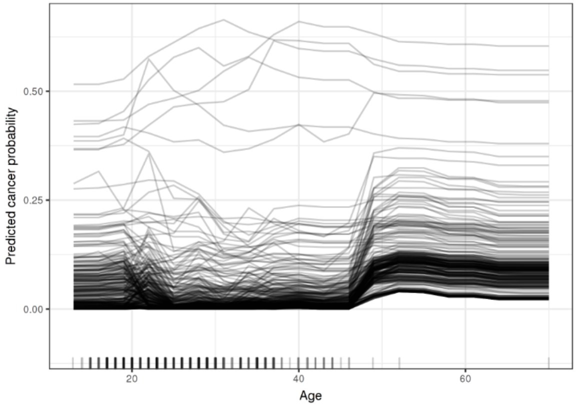
Igor Kononenko. Machine learning for medical diagnosis: history, state of the art and perspective. Artificial Intelligence in Medicine, 23:89–109, 2001.

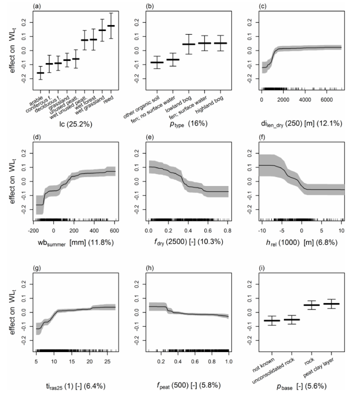
Overview of explanation in different AI fields (1)

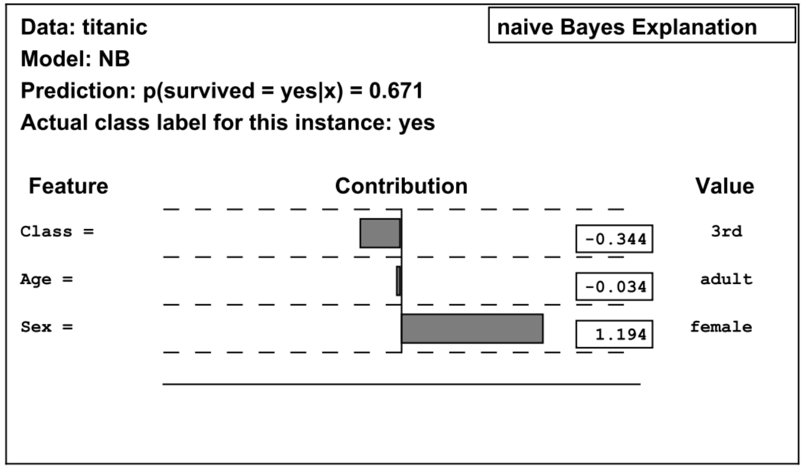
- Machine Learning (except Artificial Neural Network)

Interpretable Models:
- Linear regression,
- Logistic regression,
- Decision Tree,
- GLMs,
- GAMs
- KNNs









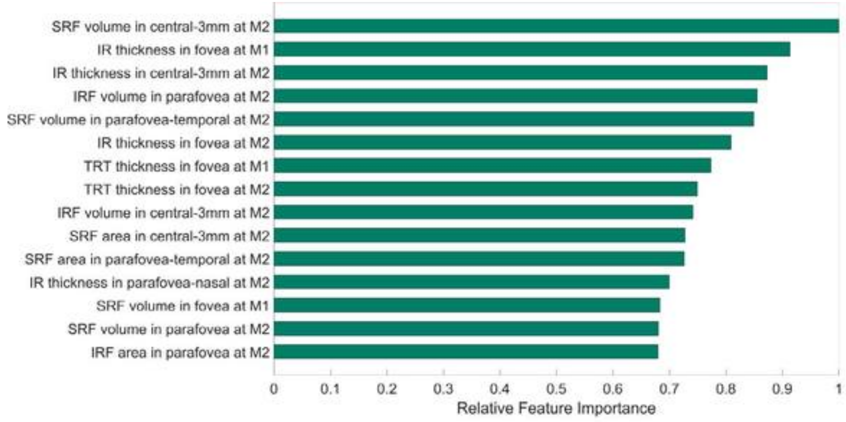
Naive Bayes model

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

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)

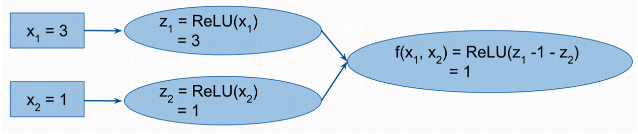


(a)

Feature Importance
Partial Dependence Plot
Individual Conditional Expectation
Sensitivity Analysis

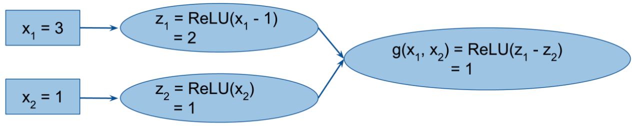
Overview of explanation in different AI fields (2)

- Machine Learning (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	$x_1 = 2,\ x_2 = -1$

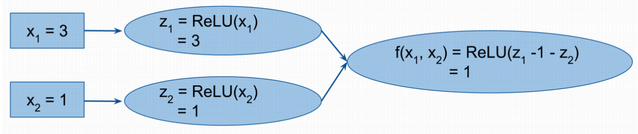
Attribution for Deep Network (Integrated gradient-based)

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

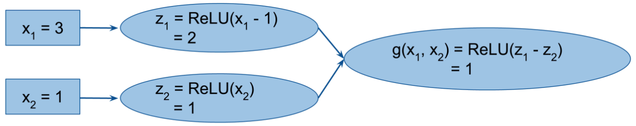
Overview of explanation in different AI fields (2)

- Machine Learning (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	$x_1 = 1.5,\ x_2 = -0.5$



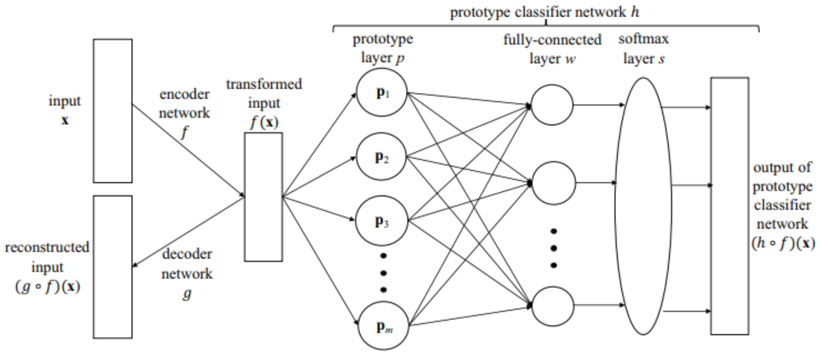
Network $g(x_1, x_2)$
Attributions at $x_1 = 3, x_2 = 1$

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

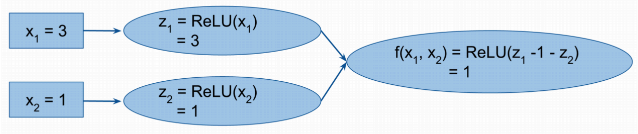


Auto-encoder

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

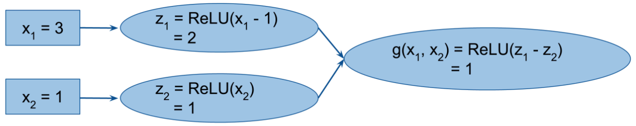
Overview of explanation in different AI fields (2)

- Machine Learning (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	$x_1 = 1.5,\ x_2 = -0.5$



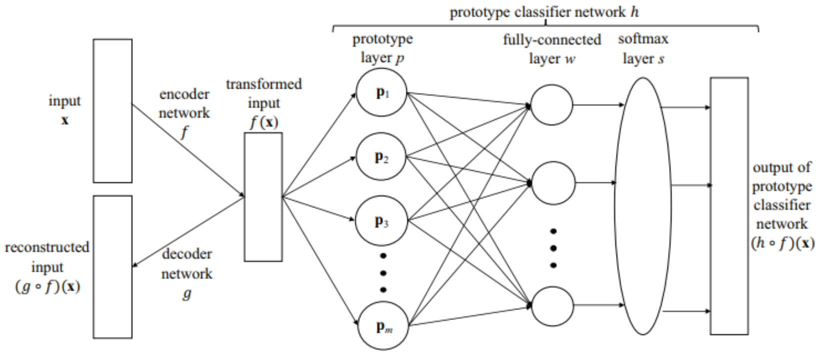
Network $g(x_1, x_2)$
Attributions at $x_1 = 3, x_2 = 1$

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)

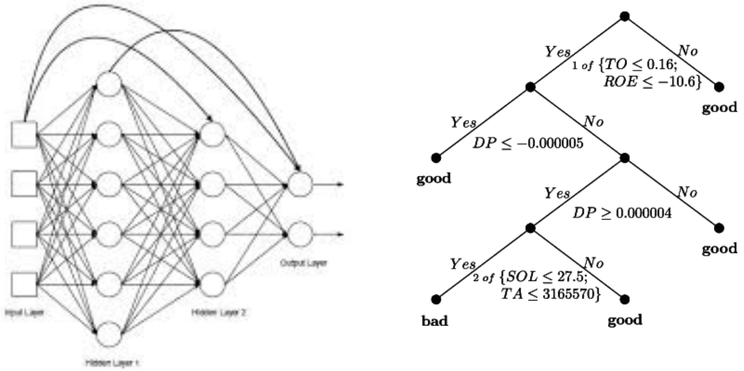
Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In ICML, pp. 3319–3328, 2017.

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



Auto-encoder

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

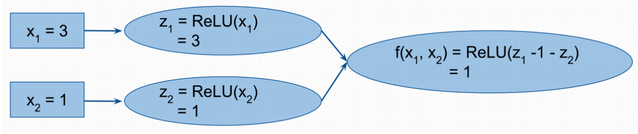


Surogate Model

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

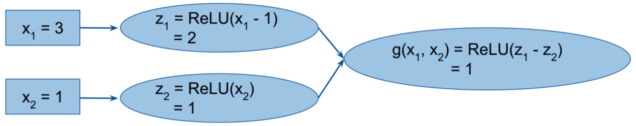
Overview of explanation in different AI fields (2)

- Machine Learning (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 $x_1 = 3, x_2 = 1$

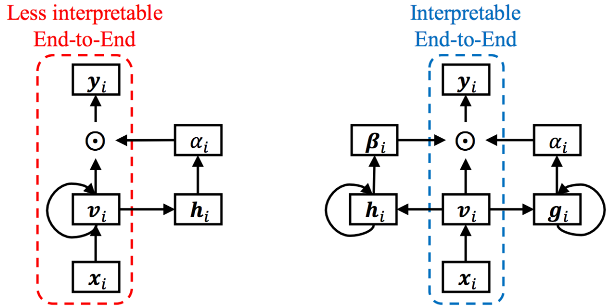
Integrated gradients	$x_1 = 1.5,\ x_2 = -0.5$
DeepLift	$x_1 = 2,\ x_2 = -1$
LRP	$x_1 = 2,\ x_2 = -1$

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

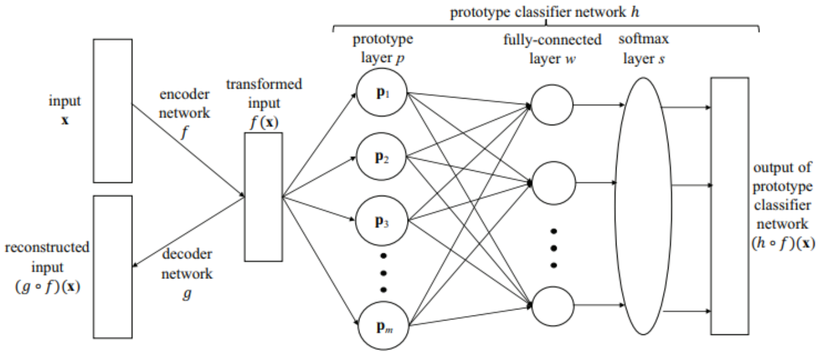
Attention Mechanism



(a) Standard attention model (b) RETAIN model

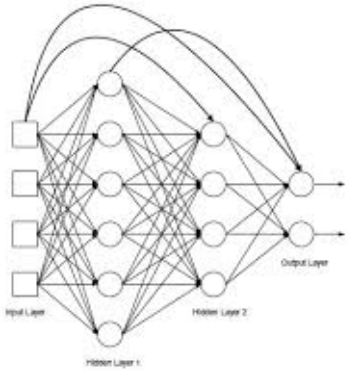
D. Bahdanau, K. Cho, and Y. Bengio. Neural machine translation by jointly learning to align and translate. International Conference on Learning Representations, 2015

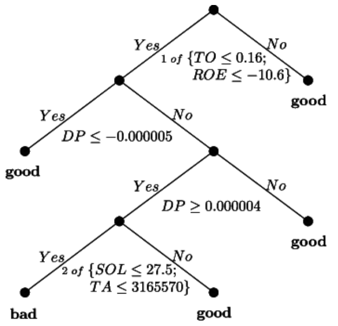
Edward Choi, Mohammad Taha Bahadori, Jimeng Sun, Joshua Kulas, Andy Schuetz, Walter F. Stewart: RETAIN: An Interpretable Predictive Model for Healthcare using Reverse Time Attention Mechanism. NIPS 2016: 3504-3512



Auto-encoder

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



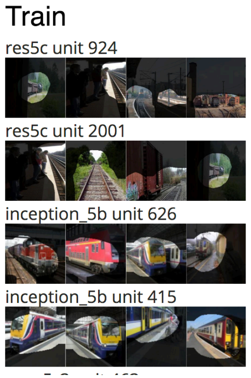


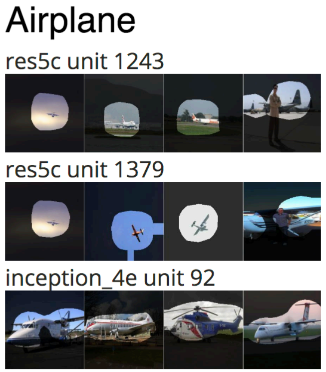
Surogate Model

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

Overview of explanation in different AI fields (3)

- Computer Vision



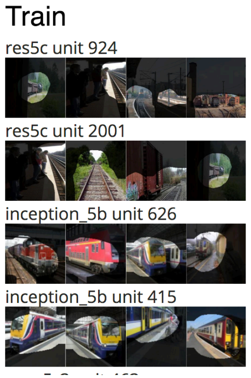


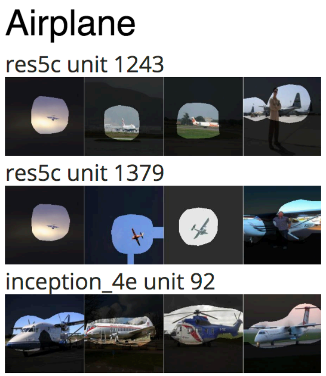
Interpretable Units

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 different AI fields (3)

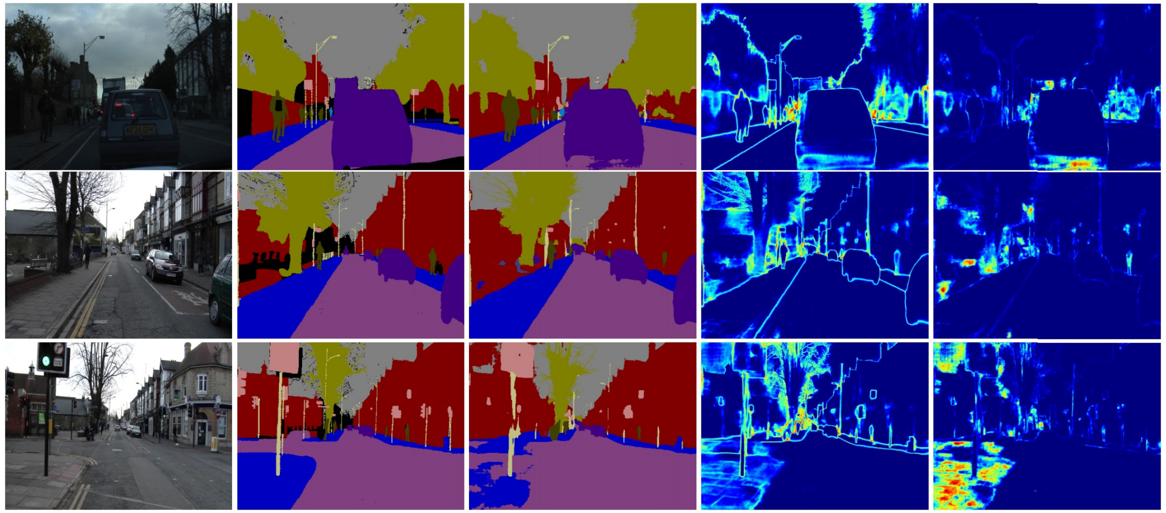
- Computer Vision





Interpretable Units

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



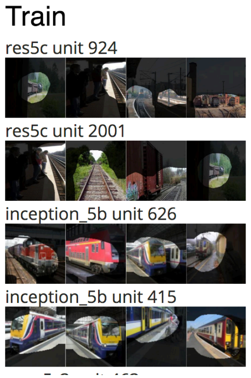
(a) Input Image (b) Ground Truth (c) Semantic Segmentation (d) Aleatoric Uncertainty (e) Epistemic Uncertainty

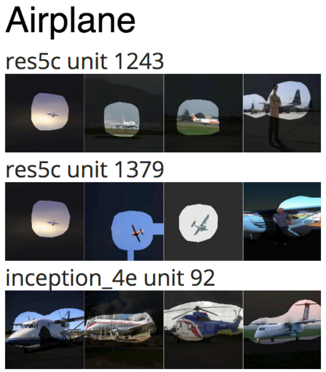
Uncertainty Map

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

Overview of explanation in different AI fields (3)

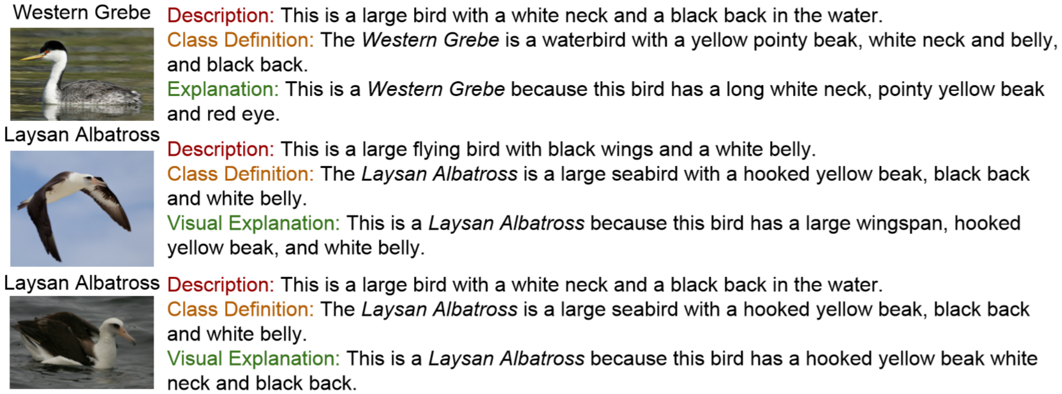
- Computer Vision





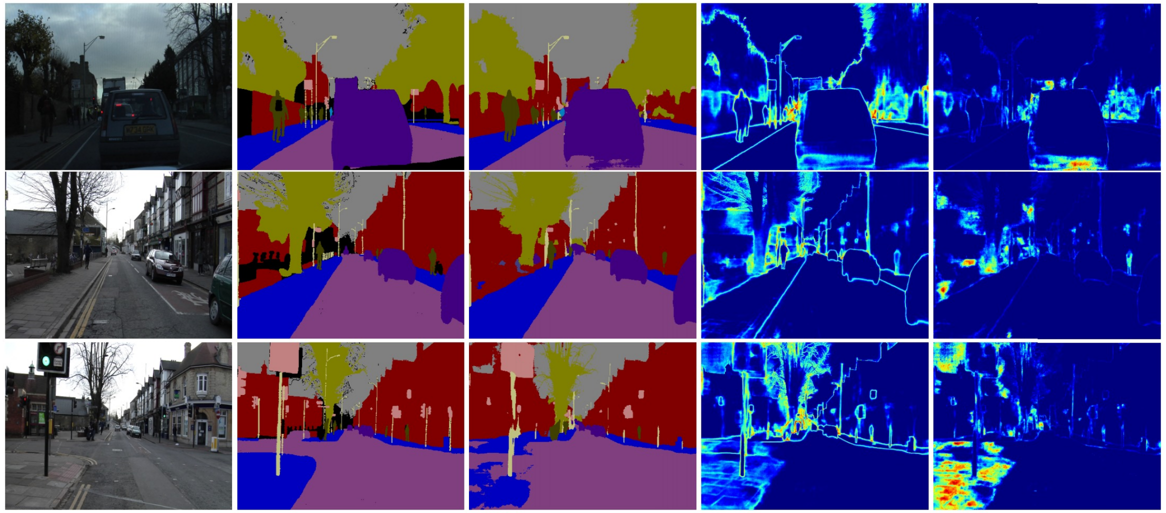
Interpretable Units

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



Visual Explanation

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



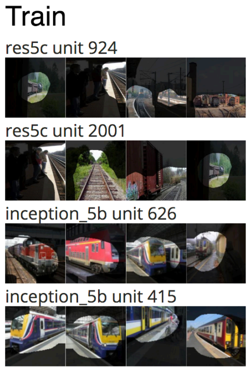
(a) Input Image (b) Ground Truth (c) Semantic Segmentation (d) Aleatoric Uncertainty (e) Epistemic Uncertainty

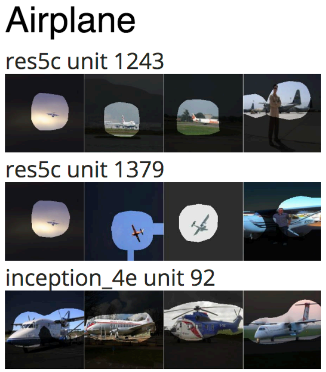
Uncertainty Map

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

Overview of explanation in different AI fields (3)

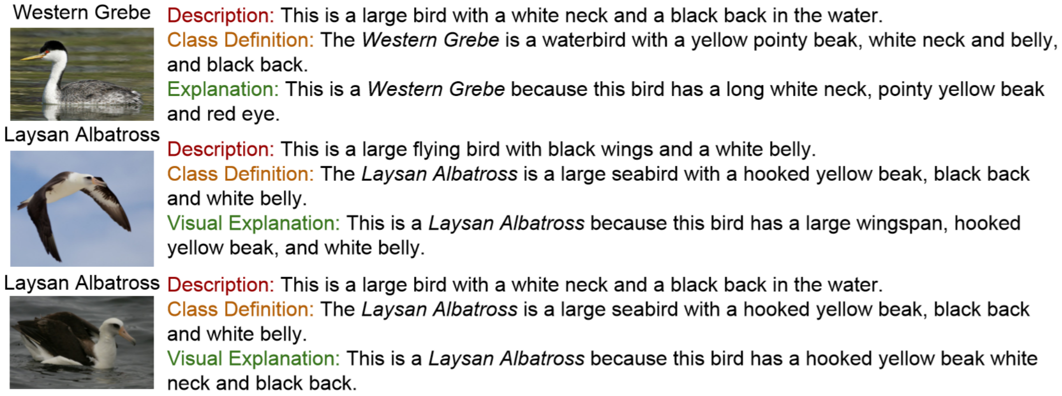
- Computer Vision





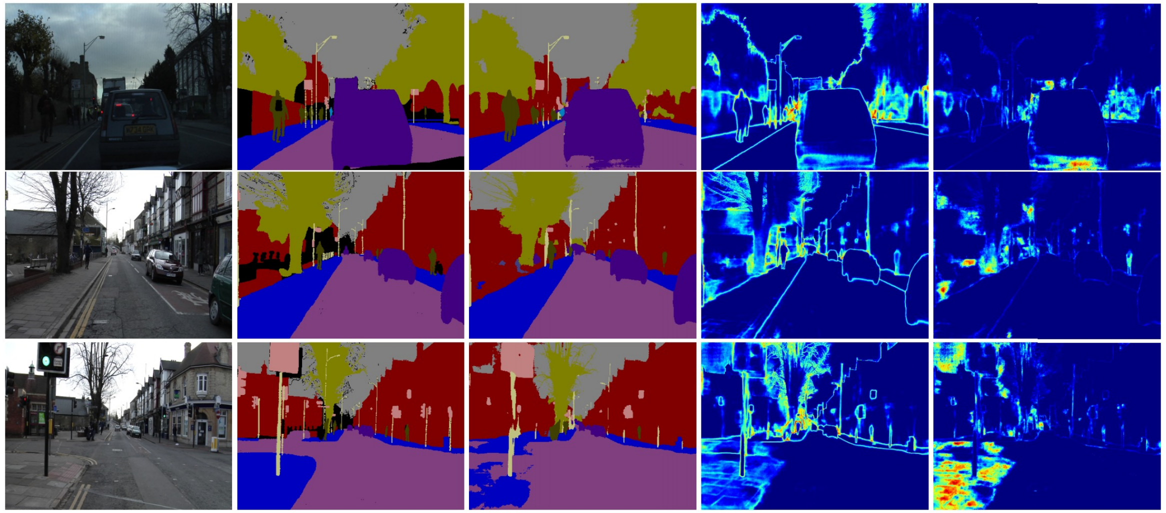
Interpretable Units

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



Visual Explanation

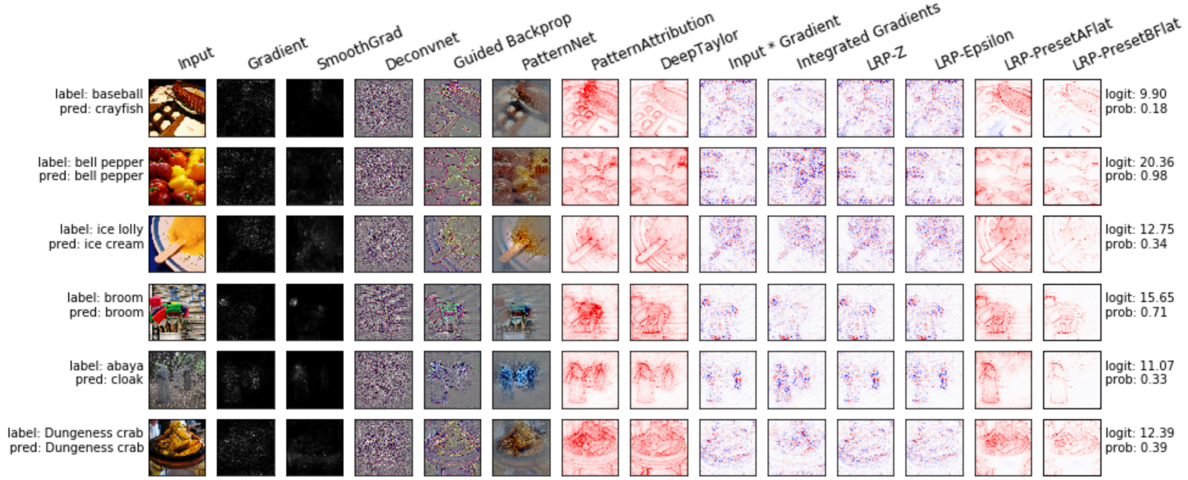
Lisa Anne Hendricks, Zeynep Akata, Marcus Rohrbach, Jeff Donahue, Bernt Schiele, Trevor Darrell: Generating Visual Explanations. ECCV (4) 2016: 3-19



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

Uncertainty Map

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

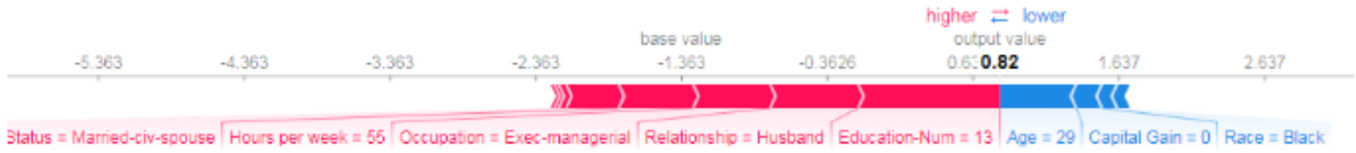


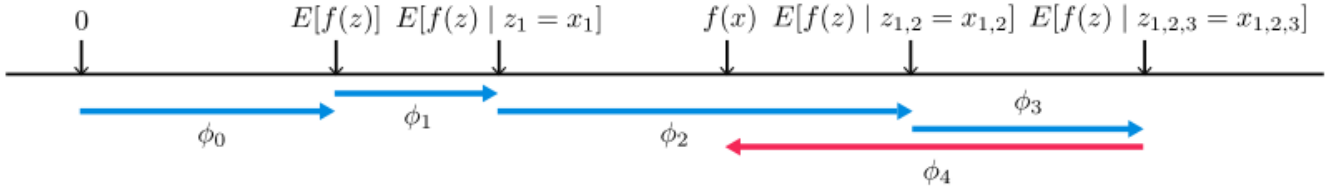
Saliency Map

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

Overview of explanation in different AI fields (4)

- Game Theory



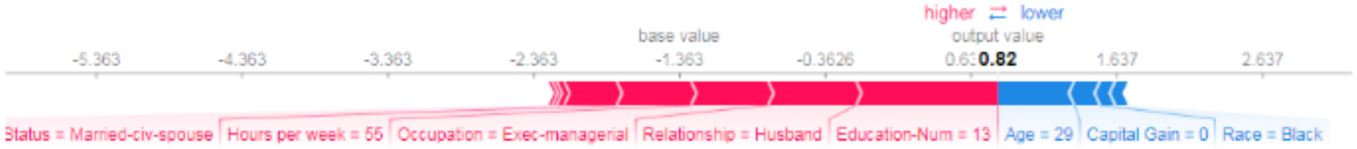


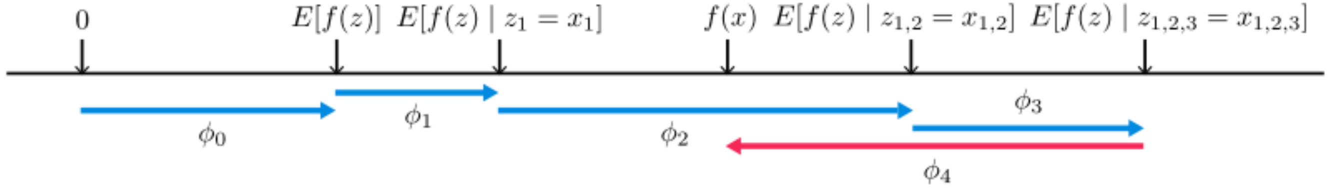
Shapley Additive Explanation

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

Overview of explanation in different AI fields (4)

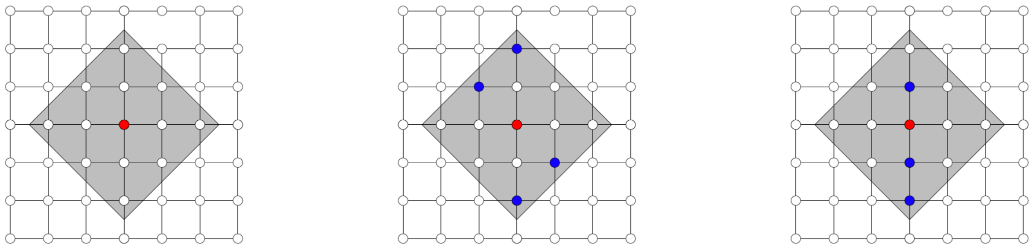
- Game Theory





Shapley Additive Explanation

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

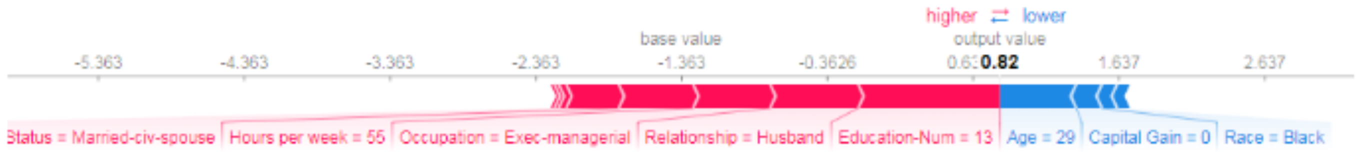


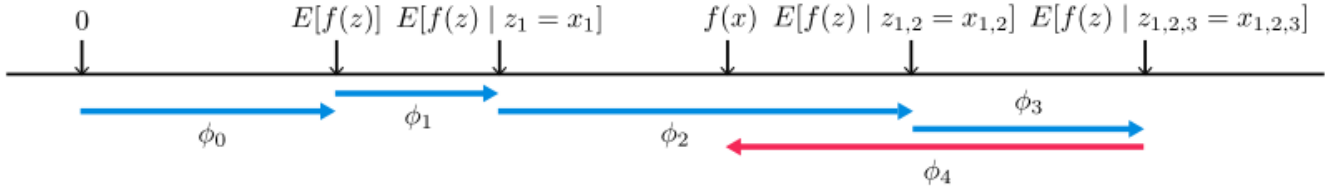
L-Shapley and C-Shapley (with graph structure)

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 (4)

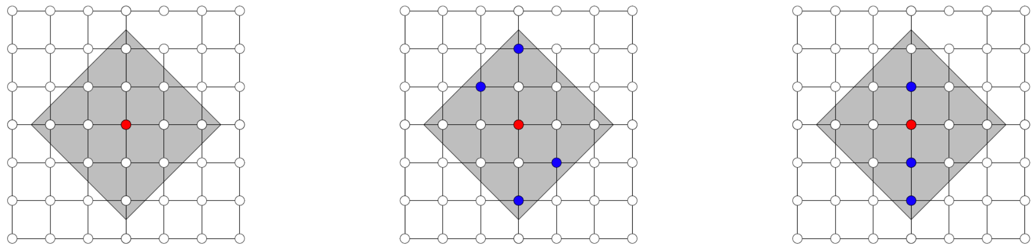
- Game Theory





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)

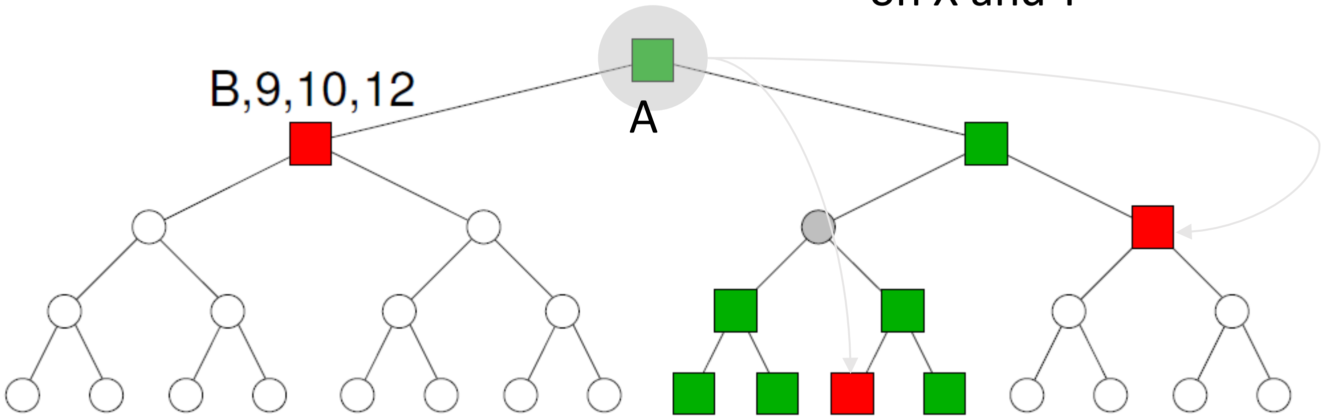
Erik Štrumbelj and Igor Kononenko. An efficient explanation of individual classifications using game theory. Journal of Machine Learning Research, 11:1–18, 2010.

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 (5)

- Search and Constraint Satisfaction

If A+1 then NEW Conflicts on X and Y



Conflicts resolution

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

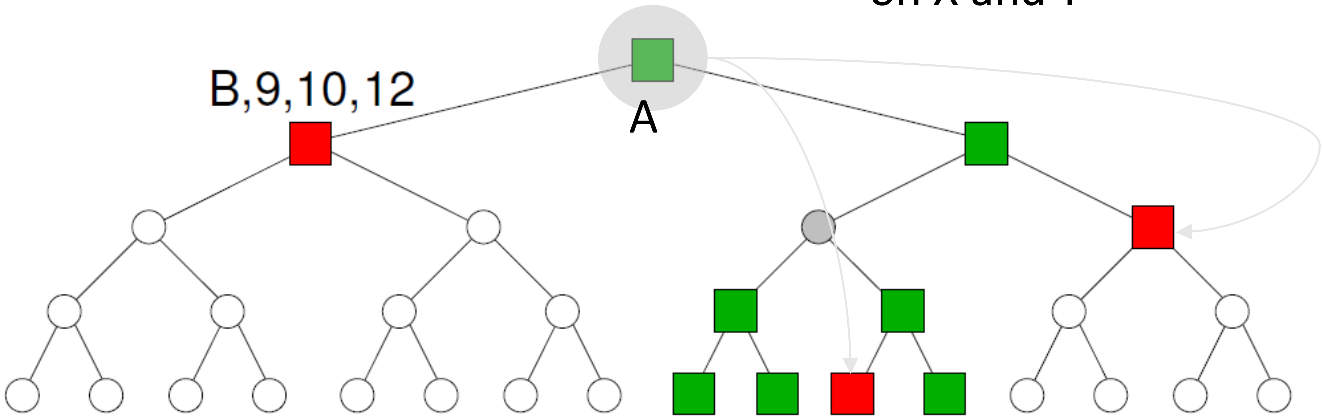
Robustness Computation

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 (5)

- Search and Constraint Satisfaction

If A+1 then NEW Conflicts on X and Y

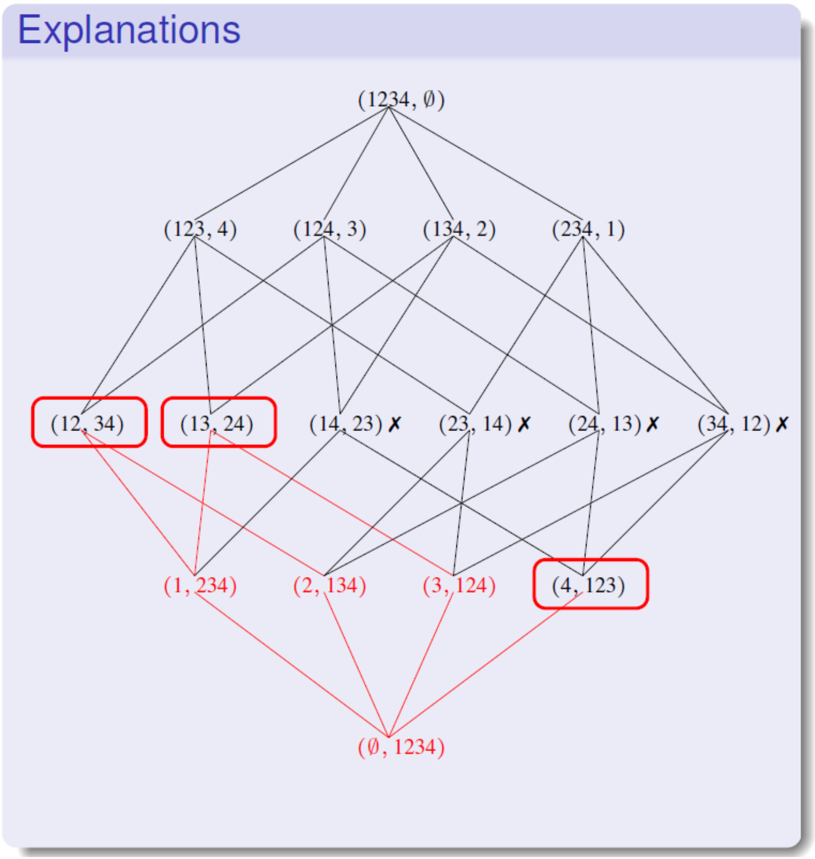


Conflicts resolution

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 (6)

Re oning

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

$\boxed{\text{A} \equiv \text{(and (at-least 3 grape) (prim GOOD WINE))}}$

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 (6)

Ref	$\vdash C \Longrightarrow C$
Trans	$\dfrac{\vdash C \Longrightarrow D,\ \vdash D \Longrightarrow E}{\vdash C \Longrightarrow E}$
Eq	$\dfrac{\vdash A \equiv B \quad \vdash C \Longrightarrow D}{\vdash C\{A/B\} \Longrightarrow D\{A/B\}}$
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	$\vdash (\text{and } (\text{all } p\ C)(\text{all } p\ D) \ldots) \equiv (\text{and } (\text{all } p\ (\text{and } C\ D)) \ldots)$

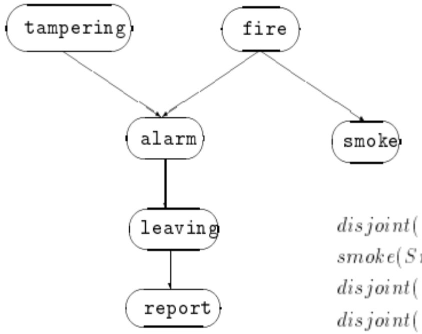
Re oning

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

$\boxed{A \equiv (\text{and (at-least 3 grape) (prim GOOD WINE)})}$

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

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



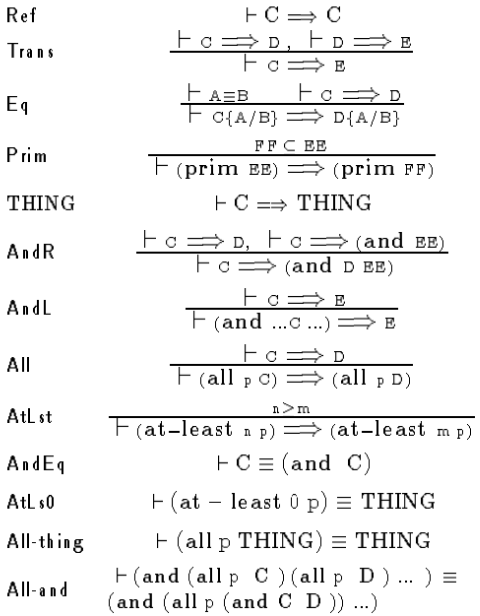
$$\begin{aligned}
P(alarm|fire \wedge \neg tampering) &= 0.99\\
P(alarm|\neg fire \wedge tampering) &= 0.85\\
P(alarm|\neg fire \wedge \neg tampering) &= 0.0001\\
P(leaving|alarm) &= 0.88\\
P(leaving|\neg alarm) &= 0.001\\
P(\text{report}|leaving) &= 0.75\\
P(\text{report}|\neg leaving) &= 0.01
\end{aligned}$$

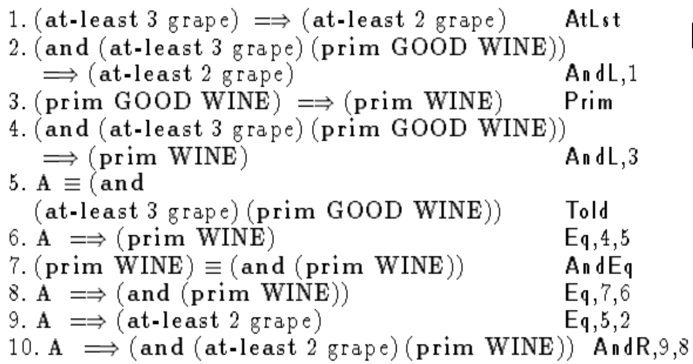
$disjoint([fire(yes) : 0.01, fire(no) : 0.99]).$
$smoke(Sm) \leftarrow fire(Fi) \wedge c\_smoke(Sm, Fi).$
$disjoint([c\_smoke(yes, yes) : 0.9, c\_smoke(no, yes) : 0.1]).$
$disjoint([c\_smoke(yes, no) : 0.01, c\_smoke(no, no) : 0.99]).$

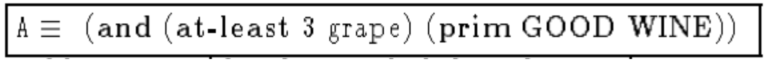
Abduction Reasoning (in Bayesian Network)

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

Overview of explanation in different AI fields (6)

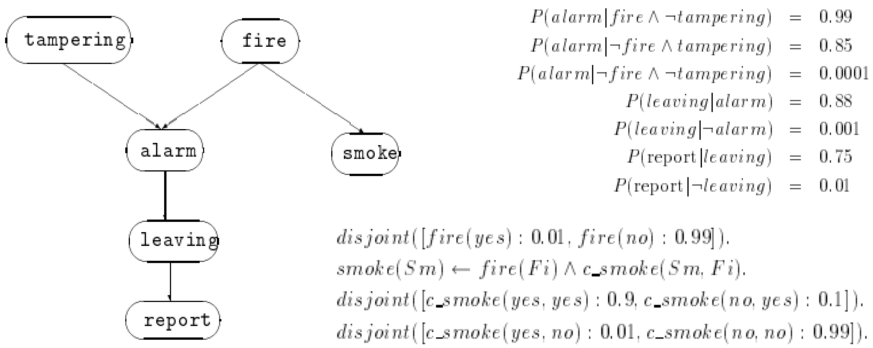






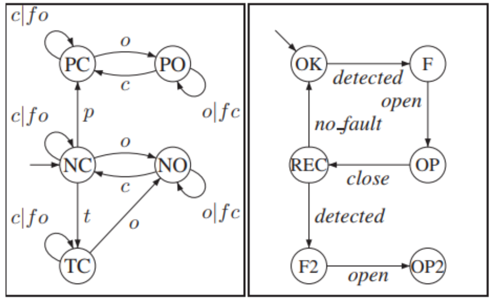
Explaining Reasoning (through Justification) e.g., Subsumption

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



Abduction Reasoning (in Bayesian Network)

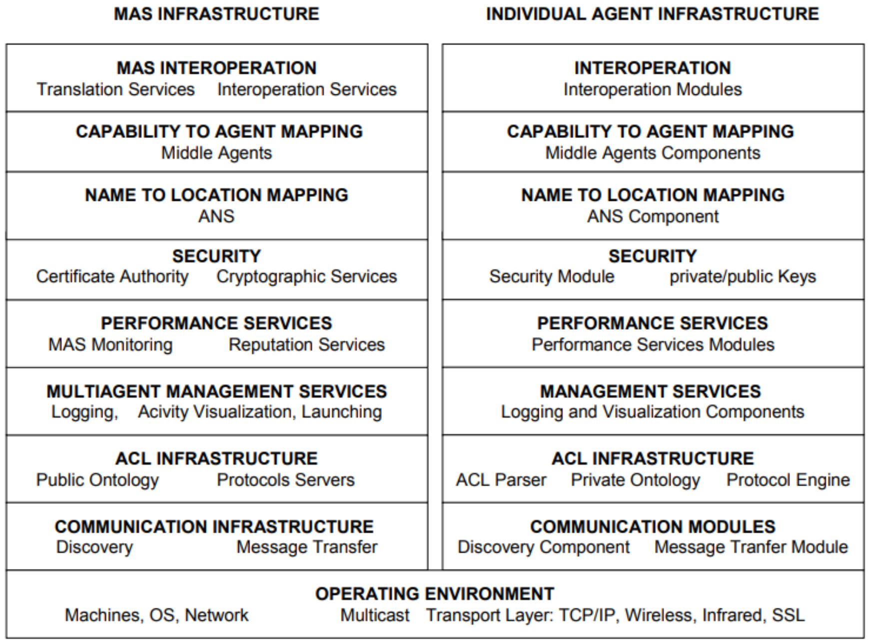
David Poole: Probabilistic Horn Abduction and Bayesian Networks. Artif. Intell. 64(1): 81-129 (1993)



Diagnosis Inference

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

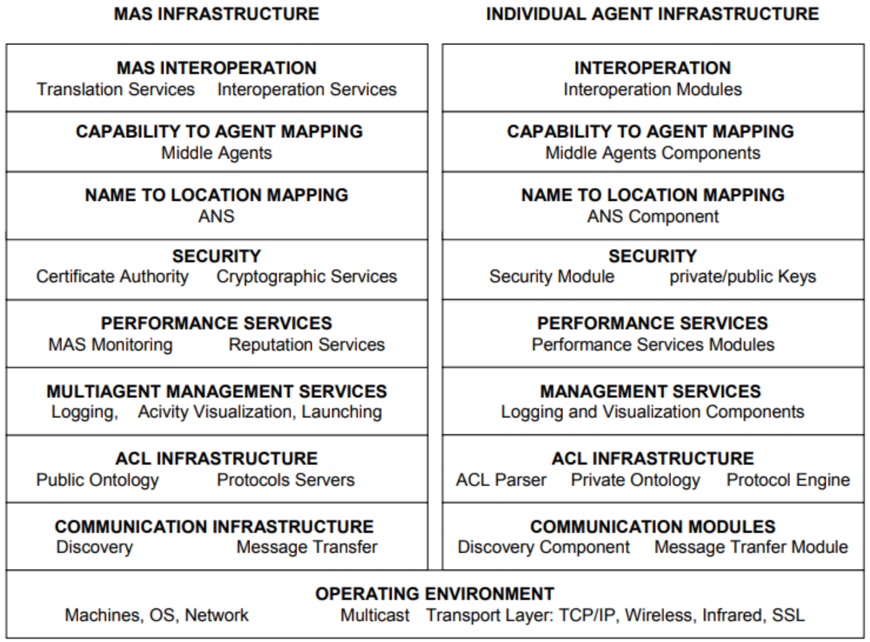
Overview of explanation in different AI fields (7)



Explanation of Agent Conflicts & Harmful Interactions

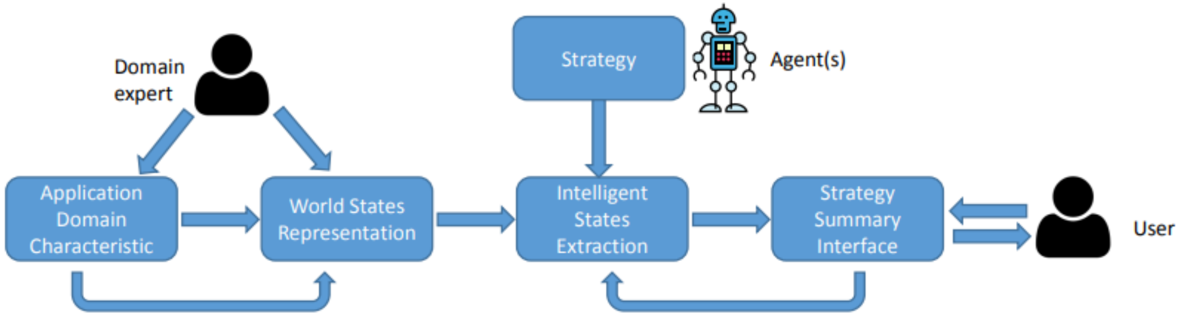
Katia P. Sycara, Massimo Paolucci, Martin Van Velsen, Joseph A. Giampapa: The RETSINA MAS Infrastructure. Autonomous Agents and Multi-Agent Systems 7(1-2): 29-48 (2003)

Overview of explanation in different AI fields (7)



Explanation of Agent Conflicts & Harmful Interactions

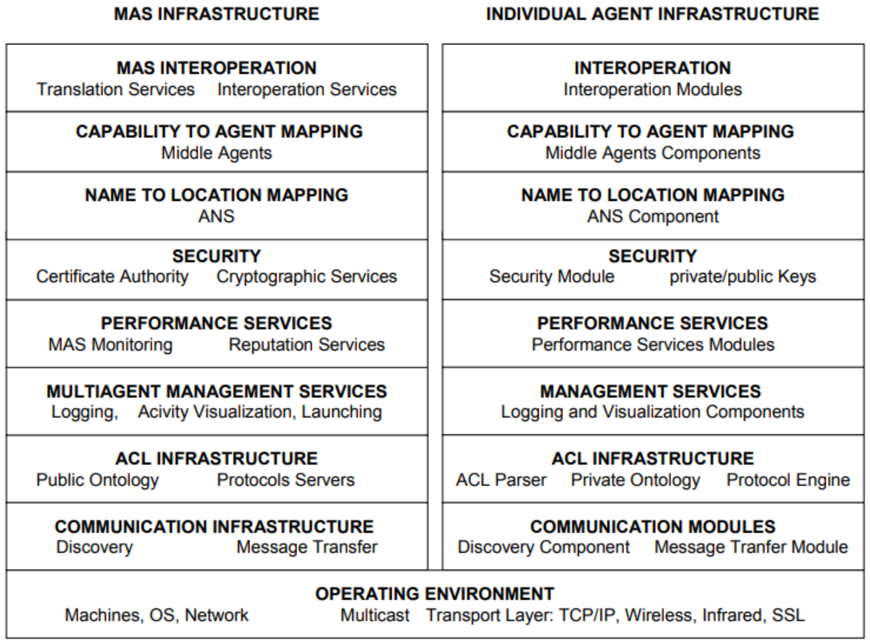
Katia P. Sycara, Massimo Paolucci, Martin Van Velsen, Joseph A. Giampapa: The RETSINA MAS Infrastructure. Autonomous Agents and Multi-Agent Systems 7(1-2): 29-48 (2003)



Agent Strategy Summarization

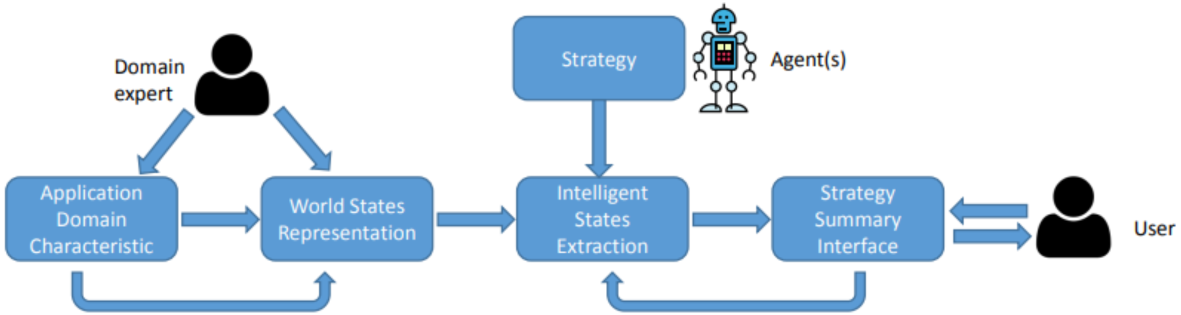
Ofra Amir, Finale Doshi-Velez, David Sarne: Agent Strategy Summarization. AAMAS 2018: 1203-1207

Overview of explanation in different AI fields (7)



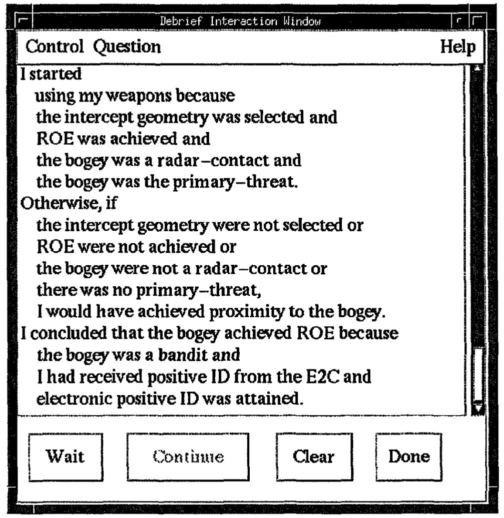
Explanation of Agent Conflicts & Harmful Interactions

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



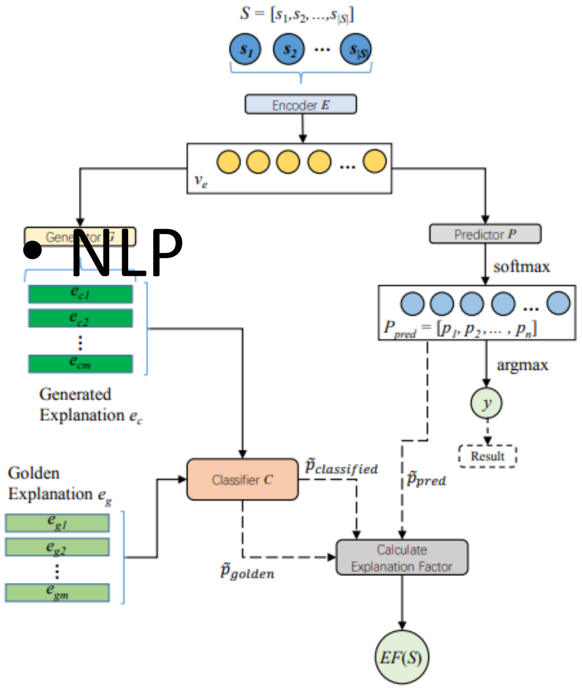
Explainable Agents

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 (8)

- 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:

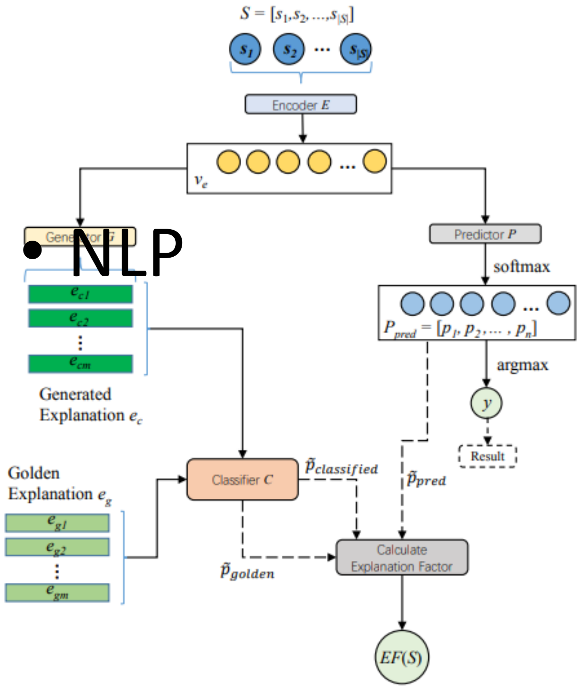
- texts in a real-world dataset;
- Numerical scores

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 (8)

- NLP

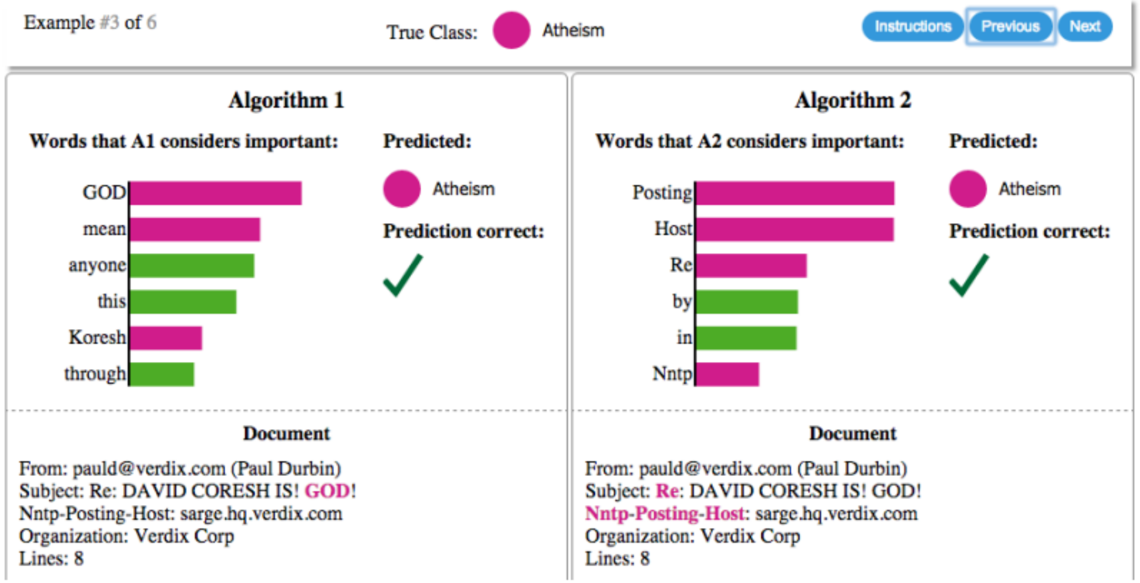


Explainable NLP

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

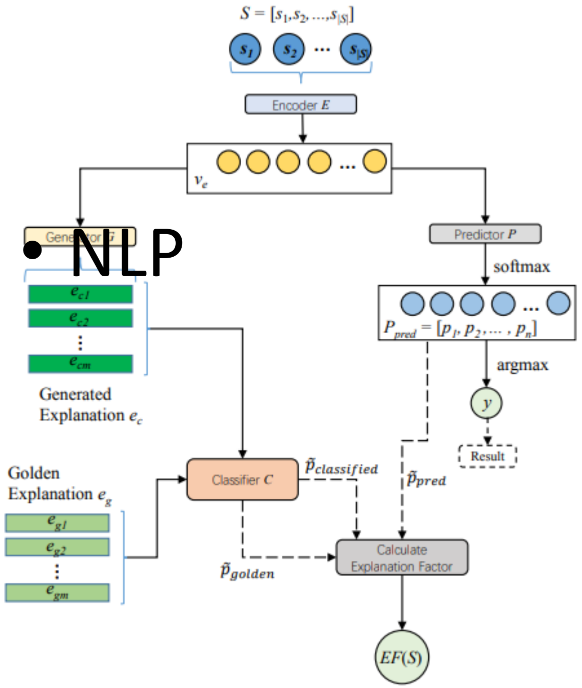
Marco Túlio Ribeiro, Sameer Singh, Carlos Guestrin: "Why Should I Trust You?": Explaining the Predictions of Any Classifier. KDD 2016: 1135-1144

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 (8)

- NLP

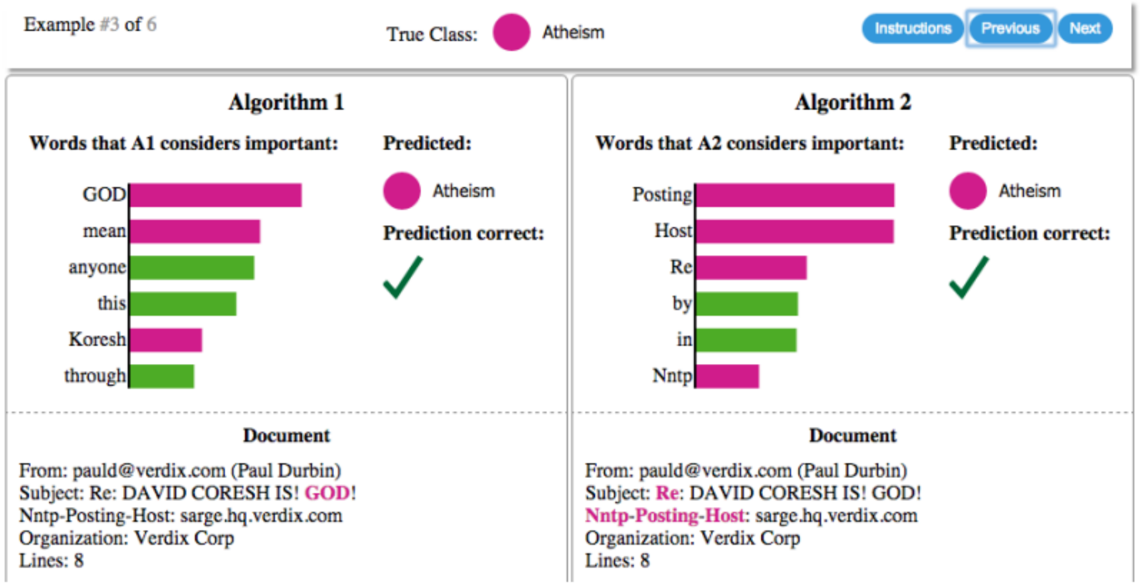


Explainable NLP

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:

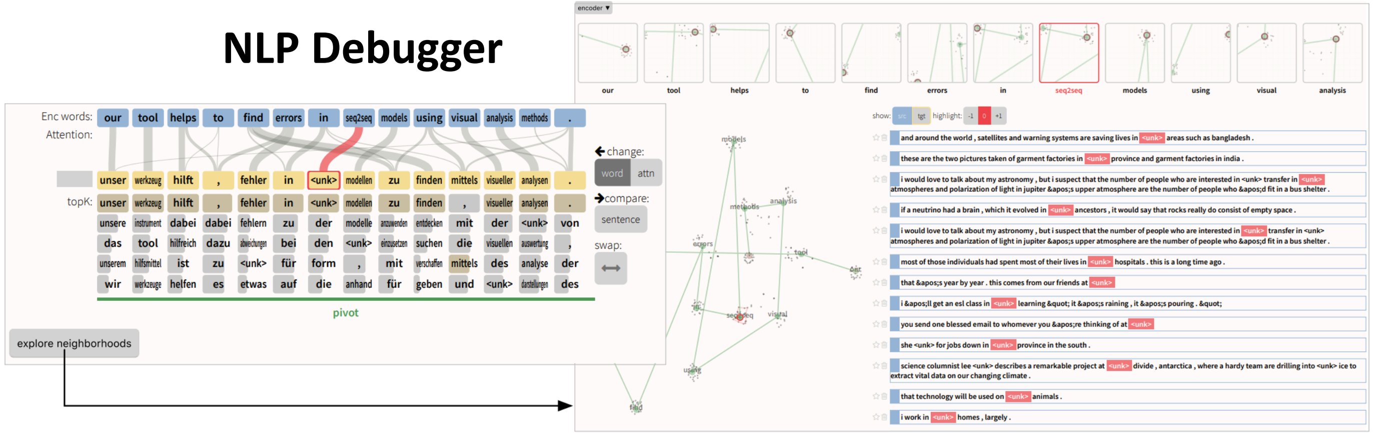
- texts in a real-world dataset;
- Numerical scores



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

NLP Debugger



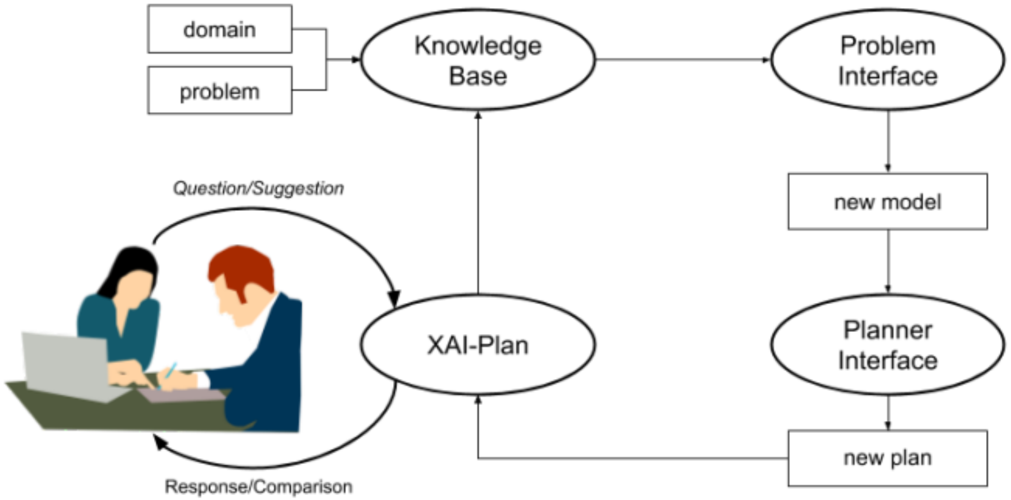
Hendrik Strobelt, Sebastian Gehrmann, Hanspeter Pfister, Alexander M. Rush: LSTMVis: A Tool for Visual Analysis of Hidden State Dynamics in Recurrent Neural Networks. IEEE Trans. Vis. Comput. Graph. 24(1): 667-676 (2018)

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 (9)

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)



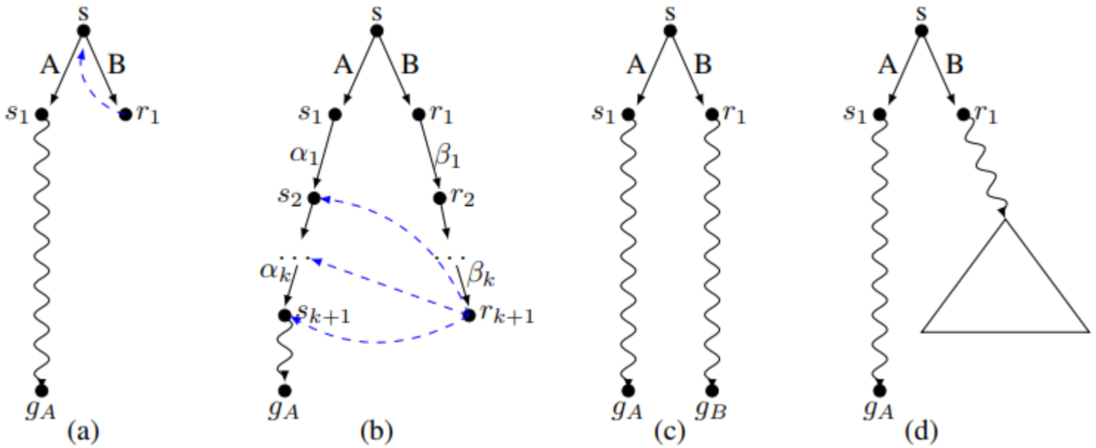
XAI Plan

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 (9)

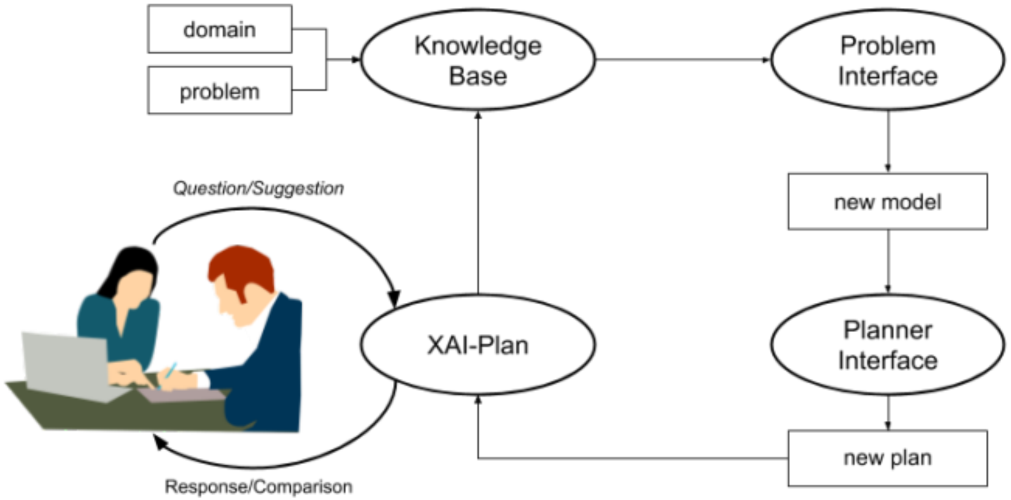
Explanation Type	R1	R2	R3	R4
Plan Patch Explanation / VAL	✗	✓	✗	✓
Model Patch Explanation	✓	✗	✓	✓
Minimally Complete Explanation	✓	✓	✗	?
Minimally Monotonic Explanation	✓	✓	✓	?
(Approximate) Minimally Complete Explanation	✗	✓	✗	✓

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



Human-in-the-loop Planning

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



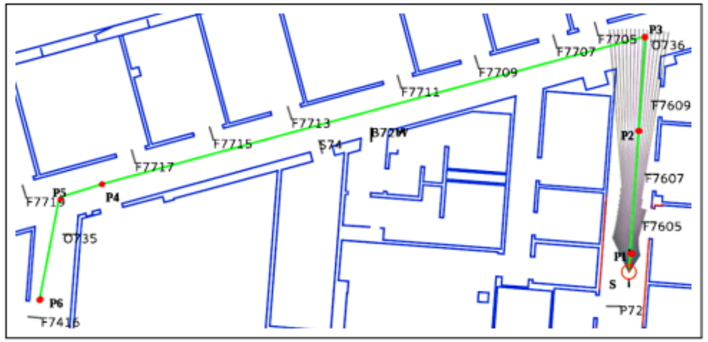
(Manual) Plan Comparison

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 (10)

- Robotics



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

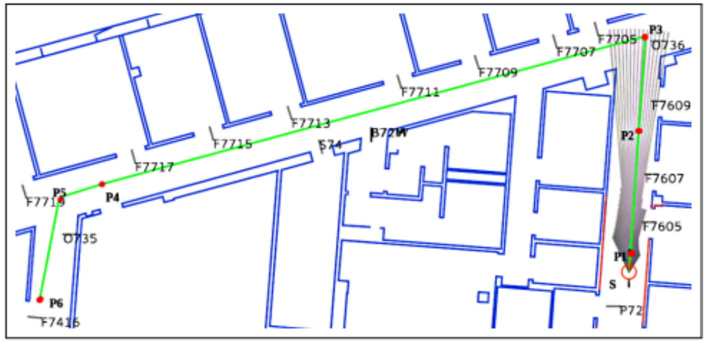
Narration of Autonomous Robot Experience

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

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

Overview of explanation in different AI fields (10)

- Robotics



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

Narration of Autonomous Robot Experience

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

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

Robot: I have decided to turn left.

Human: Why did you do that?

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

Human: How confident are you in this decision?

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

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

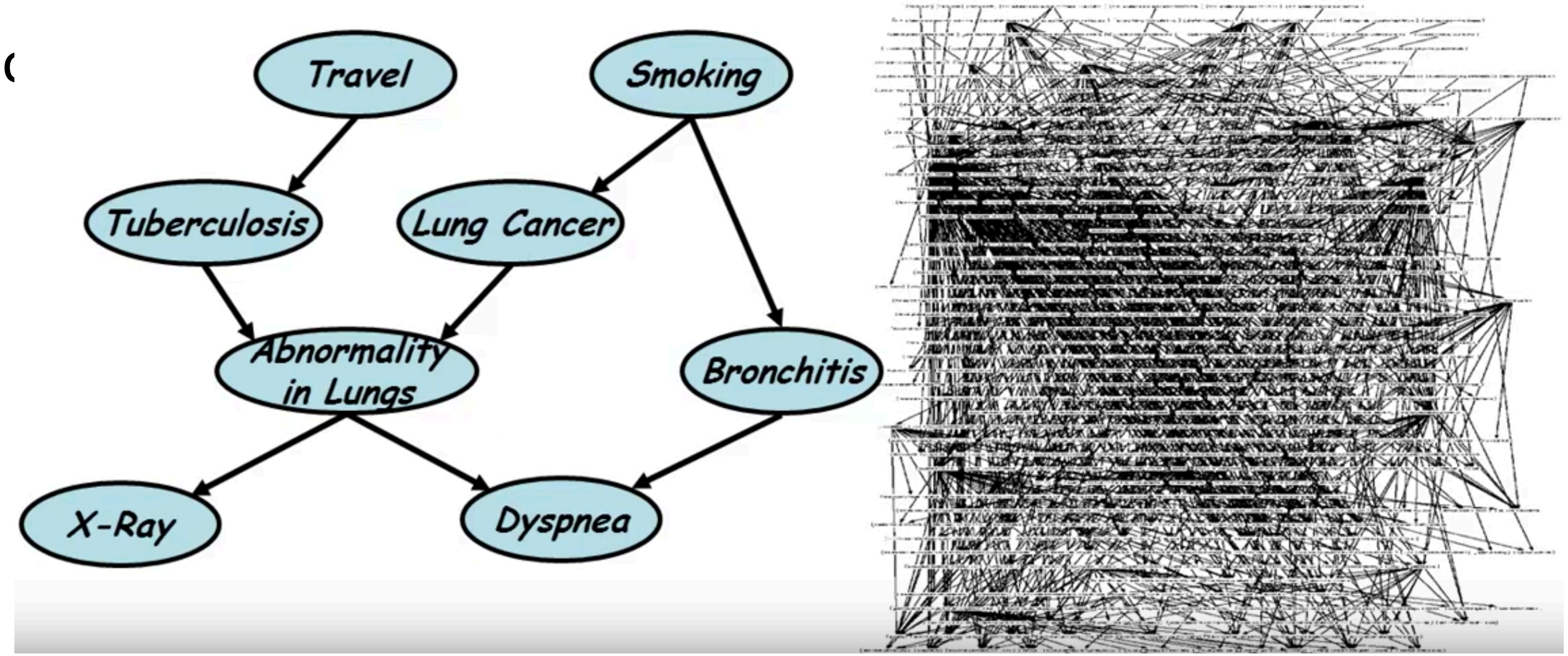
Robot: Here is the histogram of all training examples that reached this leaf. 80% of examples where this area was above 20 cm predicted the appropriate action to be "drive forward".

From Decision Tree to human-friendly information

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

Overview of explanation in different AI fields (11)

- Reaso



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

On the Role of Knowledge Graphs in Explainable AI

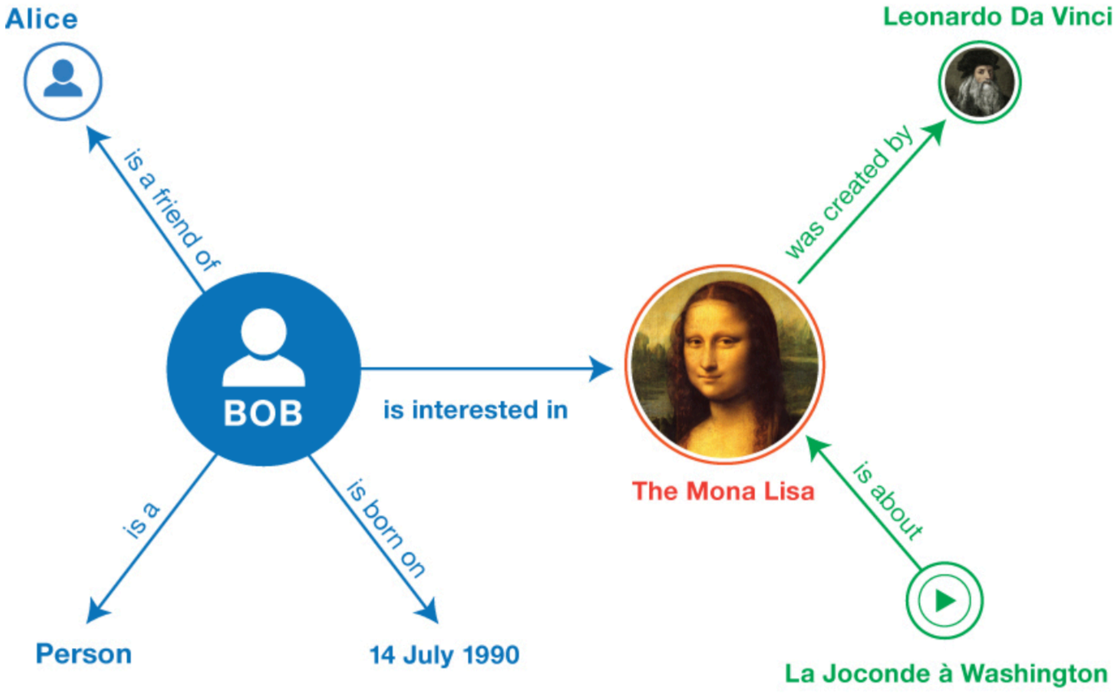
A Machine Learning Perspective

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*



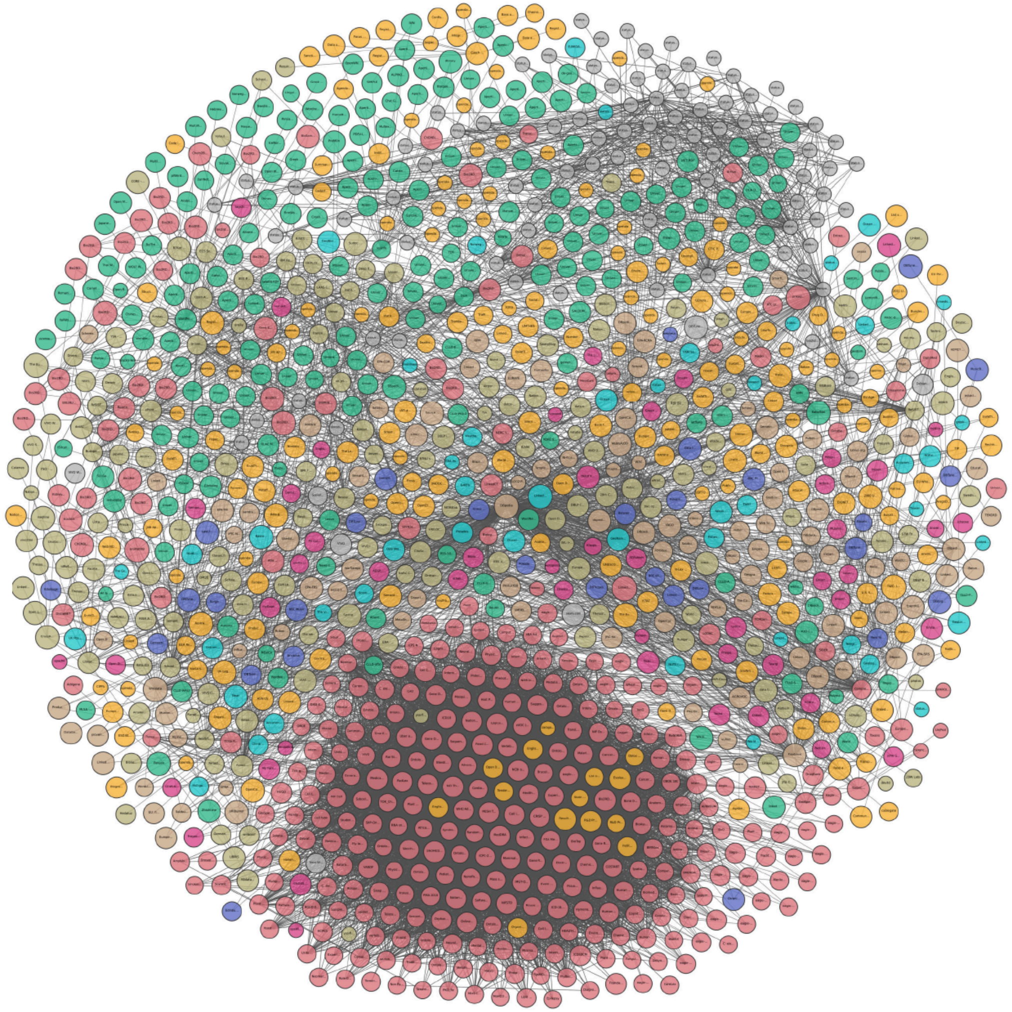
Knowledge Graph (2)

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:

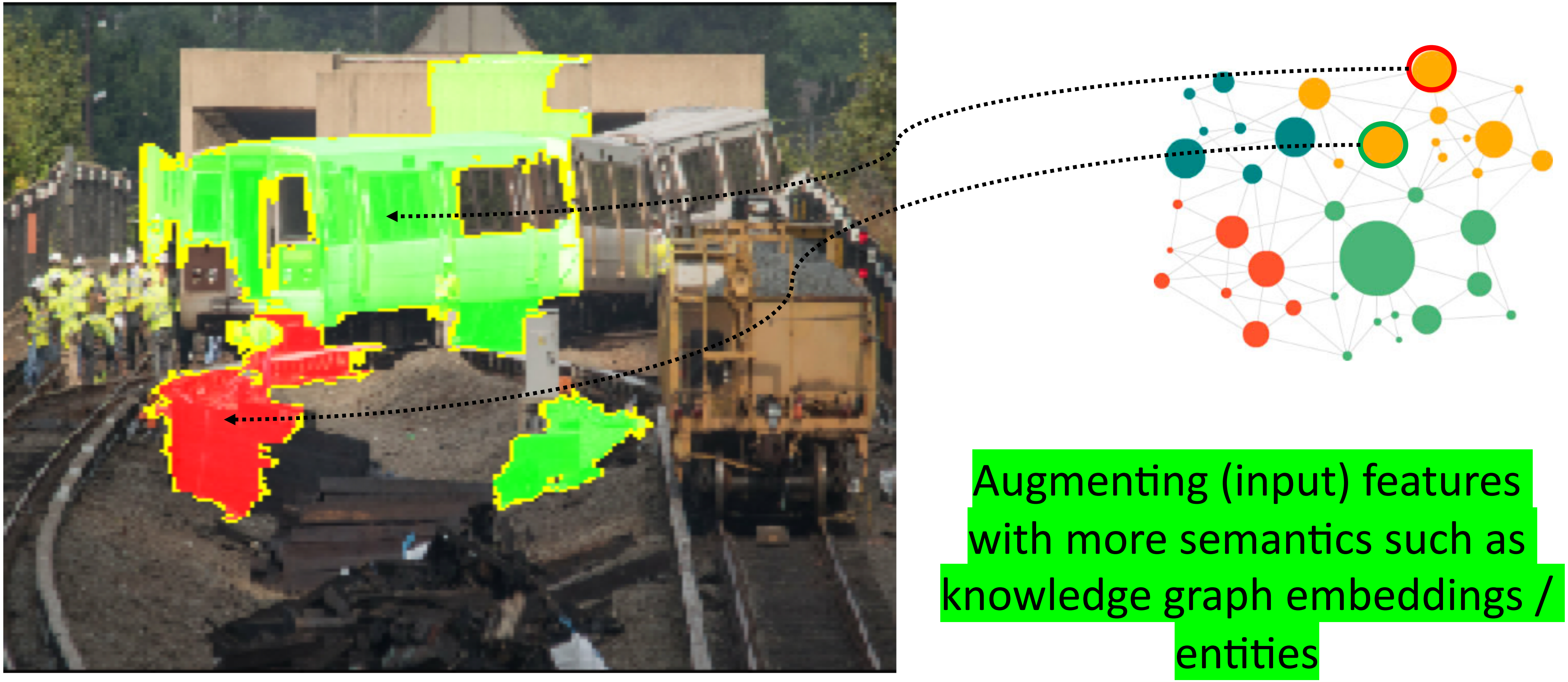
- **Manual** — curated (e.g. via experts), collaborative (e.g. via volunteers)
- **Automated** — semi-structured (e.g. from infoboxes), unstructured (e.g. from text)

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.

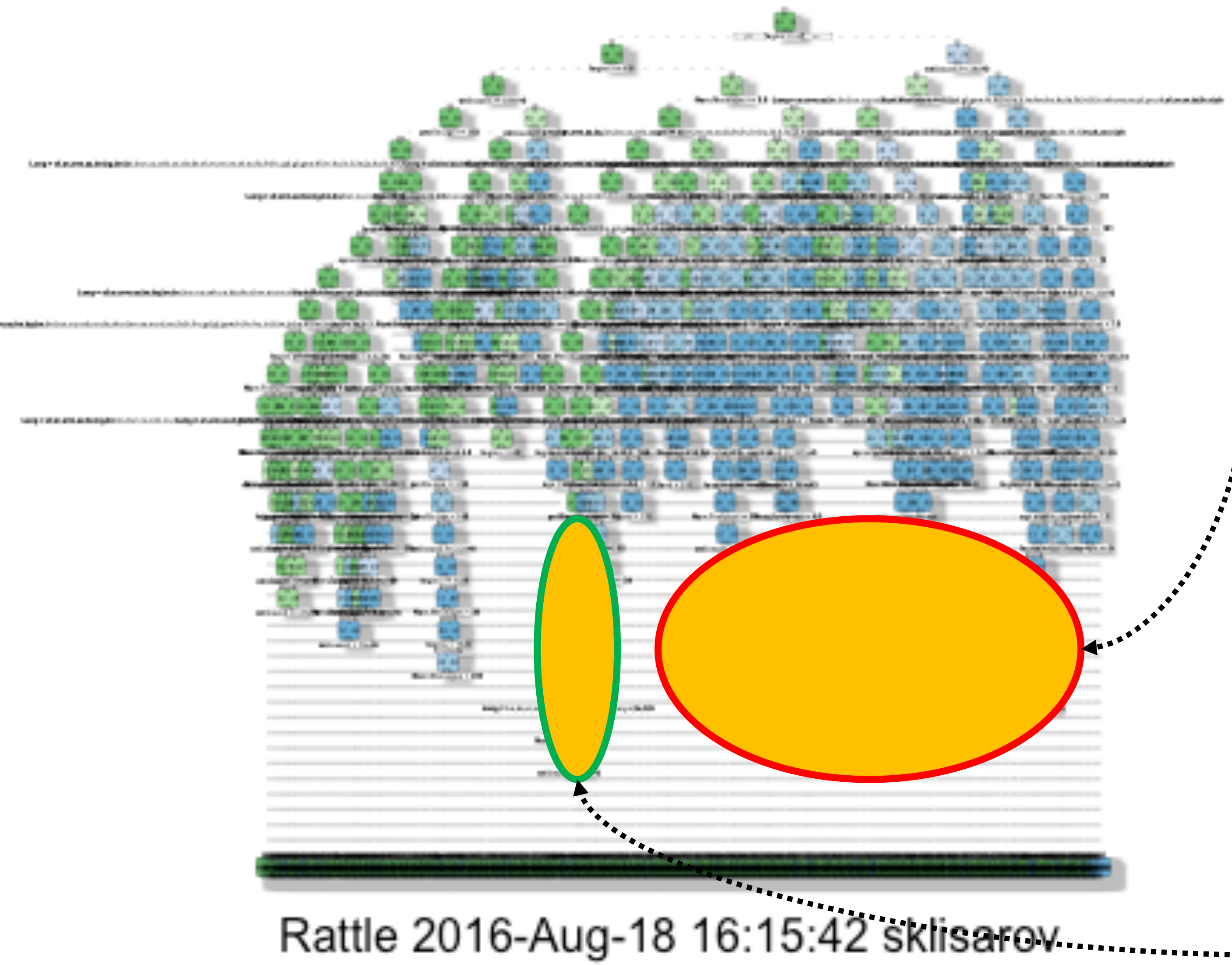
Knowledge Graph in Machine Learning (1)



Augmenting (input) features with more semantics such as knowledge graph embeddings / entities

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

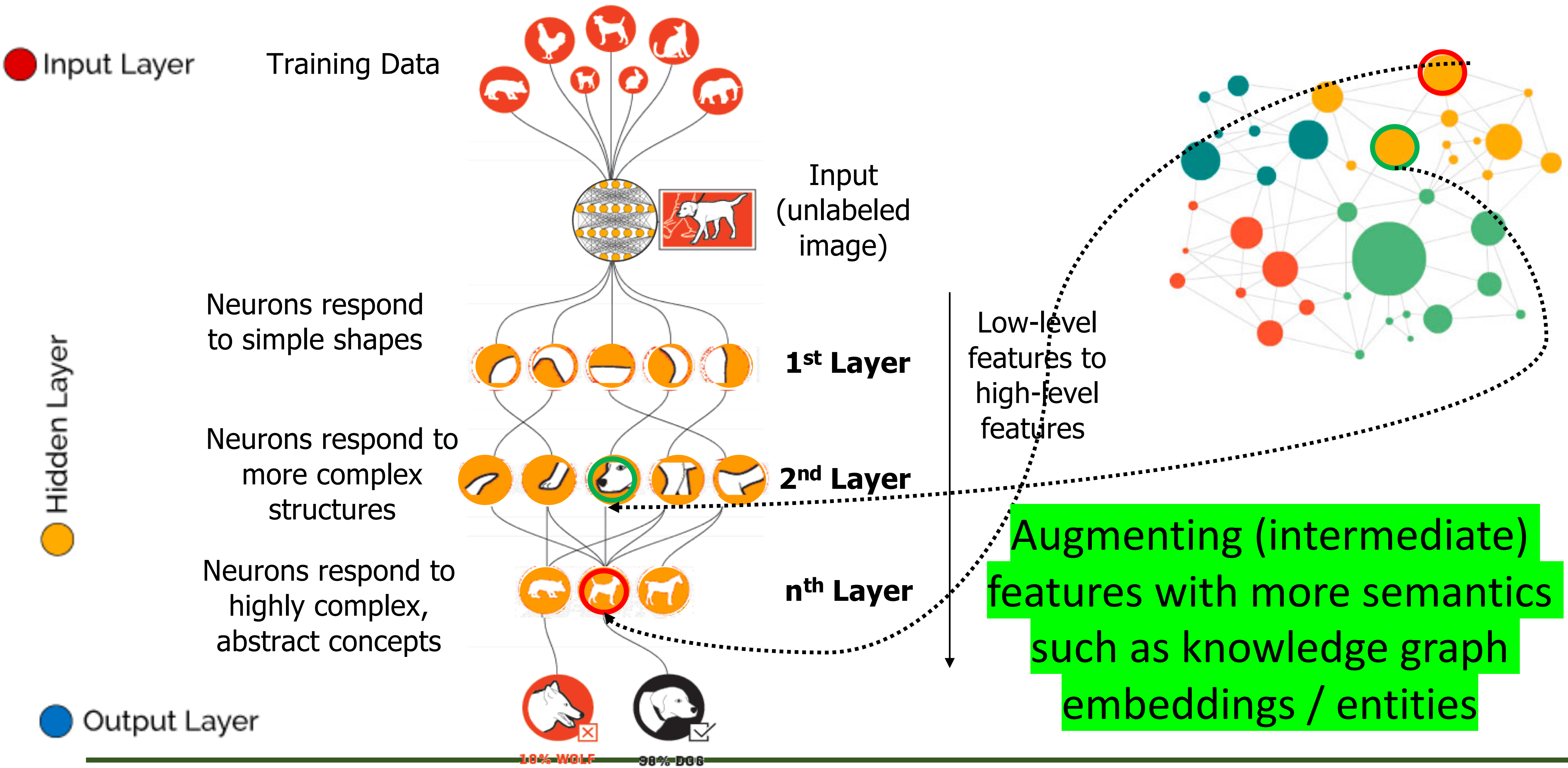
Knowledge Graph in Machine Learning (2)



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

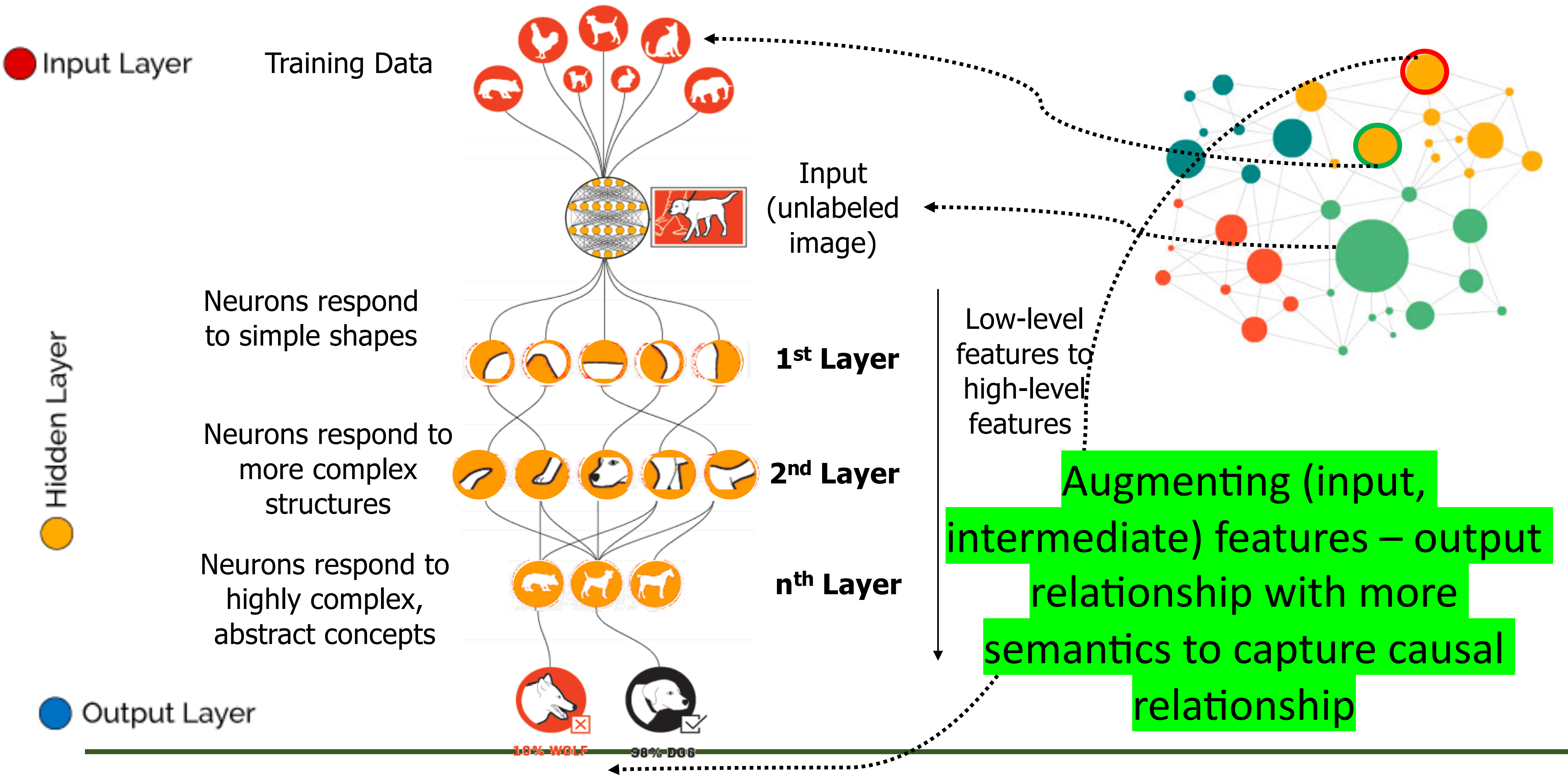
Augmenting machine learning models with more semantics such as knowledge graphs entities

Knowledge Graph in Machine Learning (3)



Augmenting (intermediate) features with more semantics such as knowledge graph embeddings / entities

Knowledge Graph in Machine Learning (4)



Augmenting (input, intermediate) features – output relationship with more semantics to capture causal relationship

Knowledge Graph in Machine Learning (5)

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

Augmenting models with semantics to support personalized explanation

Description 3: This is a public transportation accident

Knowledge Graph in Machine Learning (6)

“How to explain transfer learning with appropriate knowledge representation?

Augmenting input features and domains with semantics to support interpretable transfer learning

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

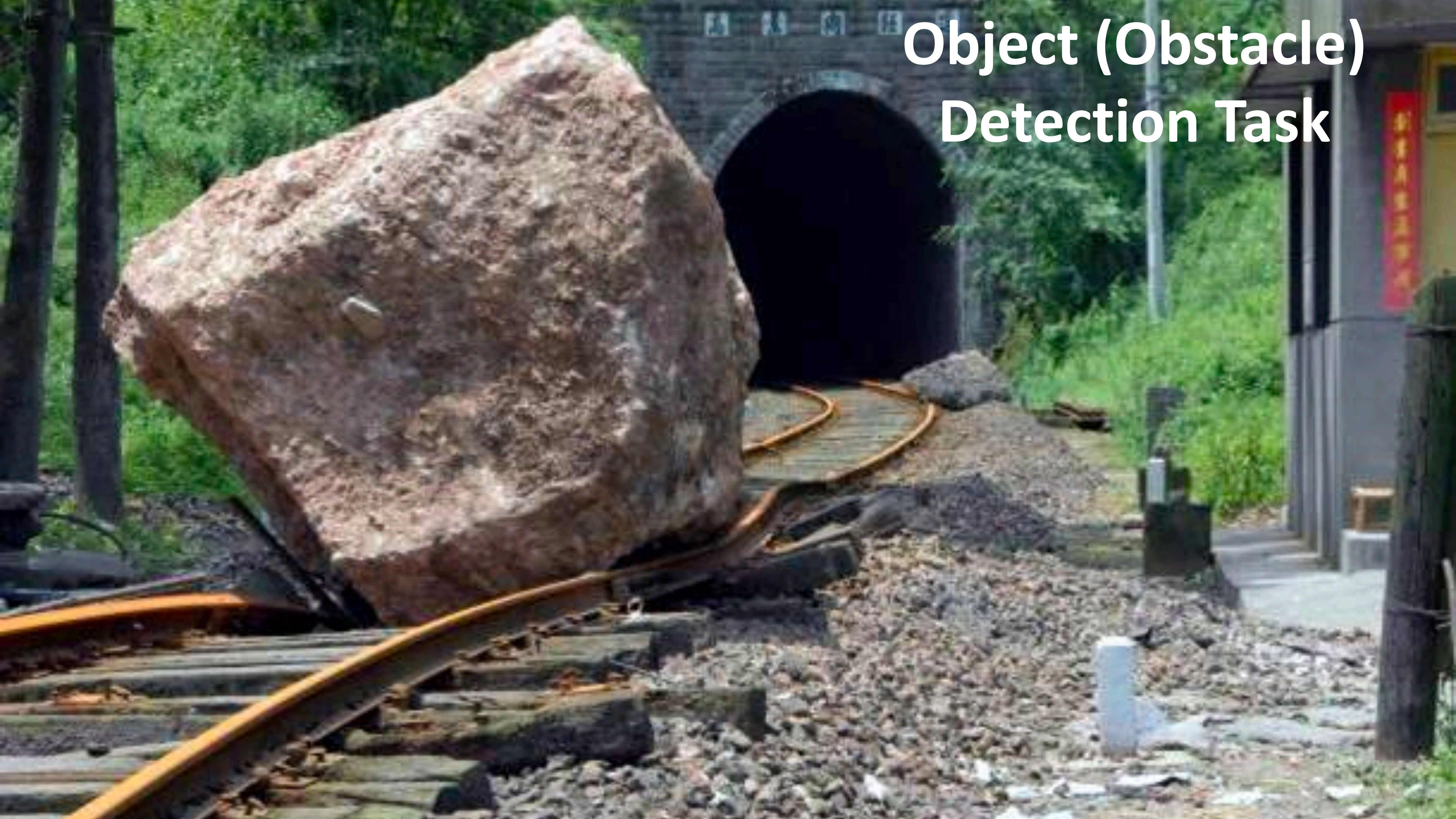
Ian Horrocks
Department of Computer Science
University of Oxford, UK

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

On One Industrial Application in Thales

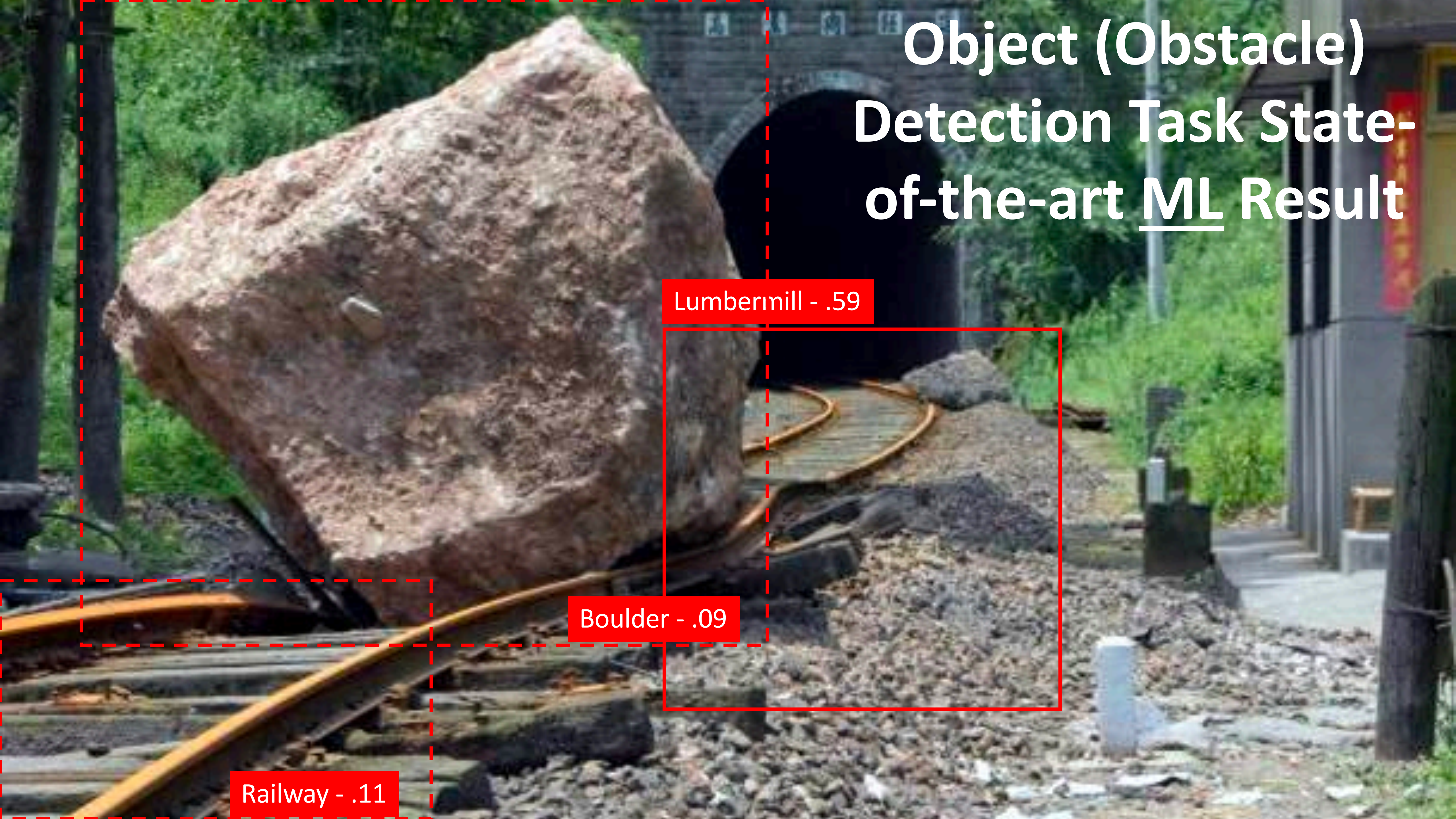
State of the Art Machine Learning Applied to Critical Systems

Object (Obstacle) Detection Task



Object (Obstacle) Detection Task State-of-the-art ML Result

Object (Obstacle) Detection Task State-of-the-art ML Result



State of the Art
XAI
Applied to Critical Systems

Object (Obstacle) Detection Task State-of-the-art XAI Result

Unfortunately, this is of NO use for a human behind the system

Let’s stay back

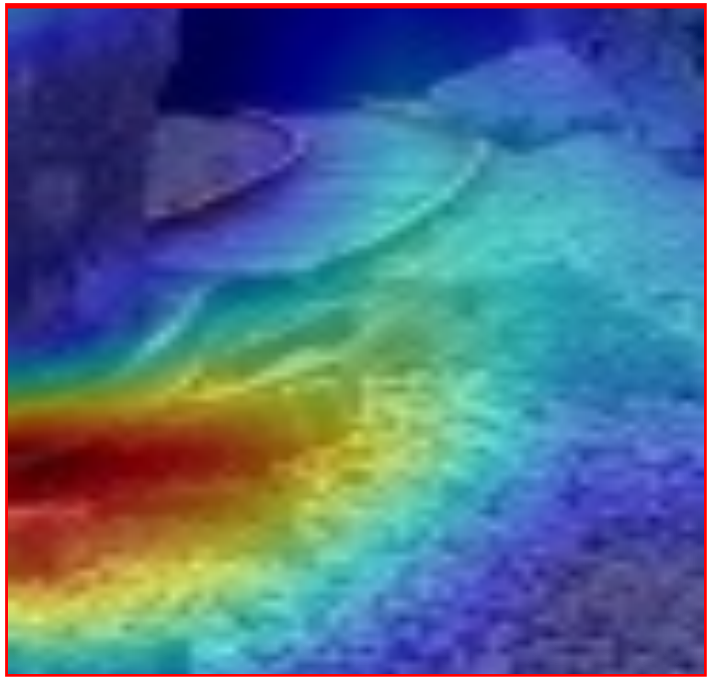
Why this Explanation? (meta explanation)

After Human Reasoning…

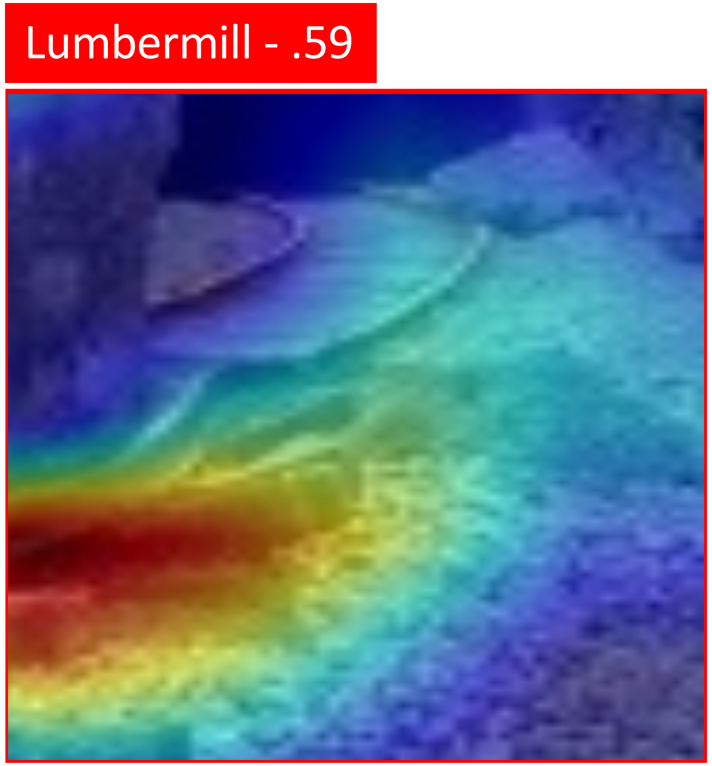
DBpedia | Browse using | Formats | Faceted Browser | Sparql Endpoint

dbo:wikiPageID	352327 (xsd:integer)
dbo:wikiPageRevisionID	734430894 (xsd:integer)
dct:subject	dbc:Sawmills dbc:Saws dbc:Ancient_Roman_technology dbc:Timber_preparation dbc:Timber_industry
http://purl.org/linguistics/gold/hypernym	dbr:Facility
rdf:type	owl:Thing dbo:ArchitecturalStructure
rdfs:comment	A sawmill or lumber mill is a facility where logs are cut into lumber. Prior to the invention of the sawmill, boards were rived (split) and planed, or more often sawn by two men with a whipsaw, one above and another in a saw pit below. The earliest known mechanical mill is the Hierapolis sawmill, a Roman water-powered stone mill at Hierapolis, Asia Minor dating back to the 3rd century AD. Other water-powered mills followed and by the 11th century they were widespread in Spain and North Africa, the Middle East and Central Asia, and in the next few centuries, spread across Europe. The circular motion of the wheel was converted to a reciprocating motion at the saw blade. Generally, only the saw was powered, and the logs had to be loaded and moved by hand. An early improvement was the developm (en)
rdfs:label	Sawmill (en)
owl:sameAs	wikidata:Sawmill dbpedia-cs:Sawmill dbpedia-de:Sawmill dbpedia-es:Sawmill

Lumbermill - .59



What is missing?



DBpedia Browse using Formats Faceted Browser Sparql Endpoint

About: Boulder

An Entity of Type : place, from Named Graph : http://dbpedia.org, within Data Space : dbpedia.org

In geology, a boulder is a rock fragment with size greater than 25.6 centimetres (10.1 in) in diameter. Smaller pieces are called cobbles and pebbles, depending on their "grain size". While a boulder may be small enough to move or roll manually, others are extremely massive. In common usage, a boulder is too large for a person to move. Smaller boulders are usually just called rocks or stones. The word boulder is short for boulder stone, from Middle English bulderston or Swedish bullersten. Boulder sized clasts are found in some sedimentary rocks, such as coarse conglomerate and boulder clay.

Property	Value
dbo:abstract	• In geology, a boulder is a rock fragment with size greater than 25.6 centimetres (10.1 in) in diameter. Smaller pieces are called cobbles and pebbles, depending on their "grain size". While a boulder may be small enough to move or roll manually, others are extremely massive. In common usage, a boulder is too large for a person to move. Smaller boulders are usually just called rocks or stones. The word boulder is short for boulder stone, from Middle English bulderston or Swedish bullersten. In places covered by ice sheets during Ice Ages, such as Scandinavia, northern North America, and Russia, glacial erratics are common. Erratics are boulders picked up by the ice sheet during its advance, and deposited during its retreat. They are called "erratic" because they typically are of a different rock type than the bedrock on which they are deposited. One of them is used as the pedestal of the Bronze Horseman in Saint Petersburg, Russia. Some noted rock formations involve giant boulders exposed by erosion, such as the Devil's Marbles in Australia's Northern Territory, the Horeke basalts in New Zealand, where an entire valley contains only boulders, and The Baths on the island of Virgin Gorda in the British Virgin Islands. Boulder sized clasts are found in some sedimentary rocks, such as coarse conglomerate and boulder clay. The climbing of large boulders is called bouldering. (en)
dbo:thumbnail	• wiki-commons:Special:FilePath/Balanced_Rock.jpg?width=300
dbo:wikiPageID	• 60784 (xsd:integer)
dbo:wikiPageRevisionID	• 743049914 (xsd:integer)
dct:subject	• dbc:Rock_formations • dbc:Rocks

DBpedia Browse using Formats Faceted Browser Sparql Endpoint

About: Rail transport

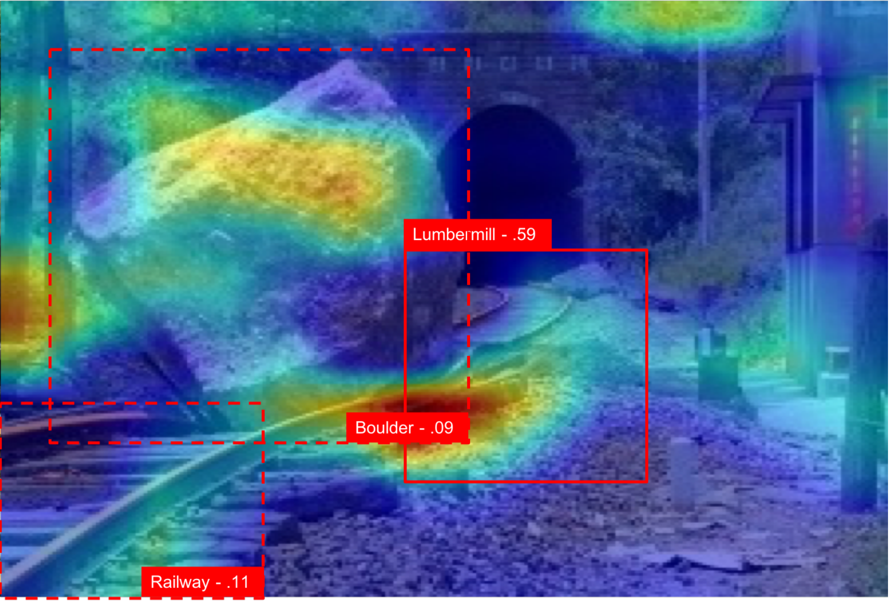
An Entity of Type : software, from Named Graph : http://dbpedia.org, within Data Space : dbpedia.org

Rail transport is a means of conveyance of passengers and goods on wheeled vehicles running on rails, also known as tracks. It is also commonly referred to as train transport. In contrast to road transport, where vehicles run on a prepared flat surface, rail vehicles (rolling stock) are directionally guided by the tracks on which they run. Tracks usually consist of steel rails, installed on ties (sleepers) and ballast, on which the rolling stock, usually fitted with metal wheels, moves. Other variations are also possible, such as slab track, where the rails are fastened to a concrete foundation resting on a prepared subsurface.

Property	Value
dbo:abstract	• Rail transport is a means of conveyance of passengers and goods on wheeled vehicles running on rails, also known as tracks. It is also commonly referred to as train transport. In contrast to road transport, where vehicles run on a prepared flat surface, rail vehicles (rolling stock) are directionally guided by the tracks on which they run. Tracks usually consist of steel rails, installed on ties (sleepers) and ballast, on which the rolling stock, usually fitted with metal wheels, moves. Other variations are also possible, such as slab track, where the rails are fastened to a concrete foundation resting on a prepared subsurface. Rolling stock in a rail transport system generally encounters lower frictional resistance than road vehicles, so passenger and freight cars (carriages and wagons) can be coupled into longer trains. The operation is carried out by a railway company, providing transport between train stations or freight customer facilities. Power is provided by locomotives which either draw electric power from a railway electrification system or produce their own power, usually by diesel engines. Most tracks are accompanied by a signalling system. Railways are a safe land transport system when compared to other forms of transport. Railway transport is capable of high levels of passenger and cargo utilization and energy efficiency, but is often less flexible and more capital-intensive than road transport, when lower traffic levels are considered. The oldest, man-hauled railways date back to the 6th century BC, with Periander, one of the Seven Sages of Greece,

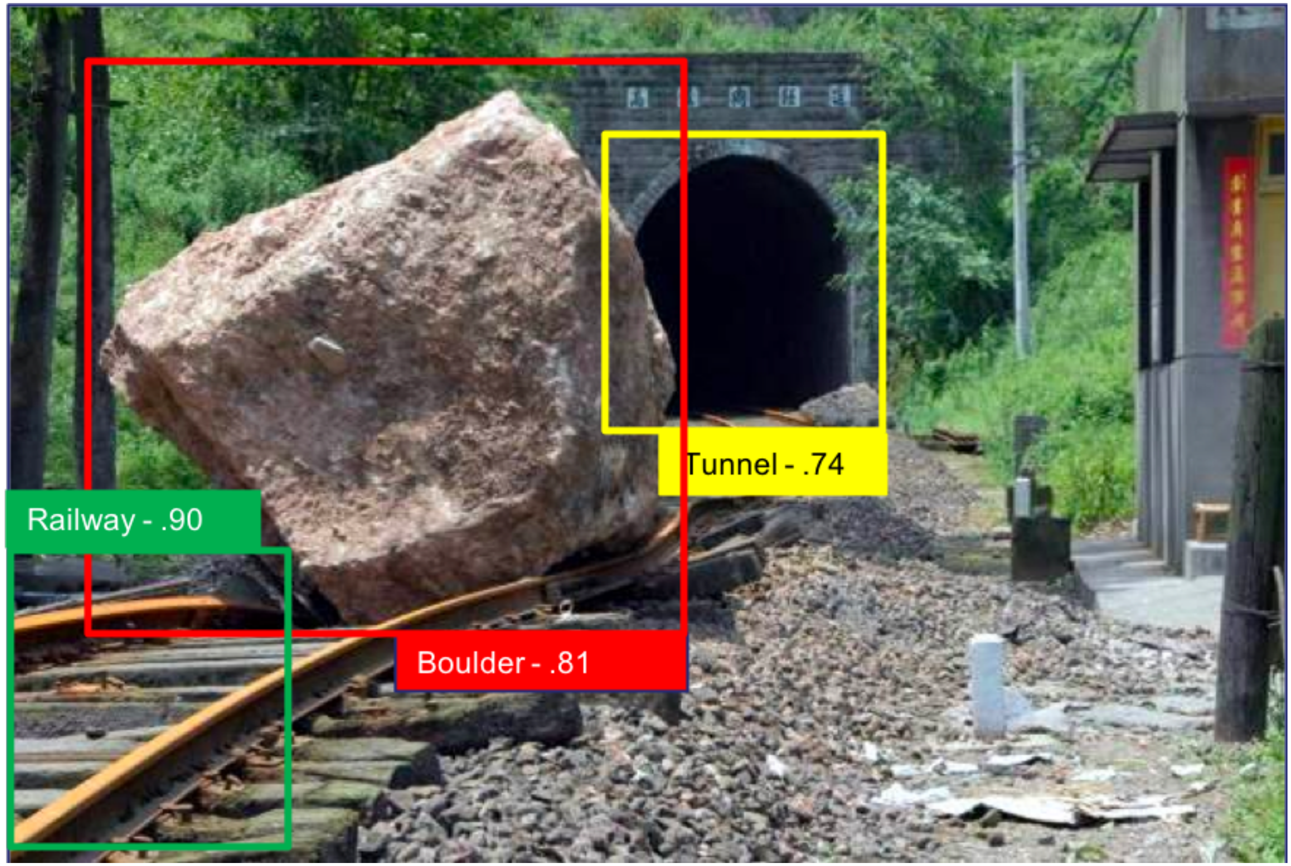
XAI Thales Platform

- **Higher accuracy with no intensive fine-tuning**
- **Human interpretable explanation**
- **Running on the edge at inference time**

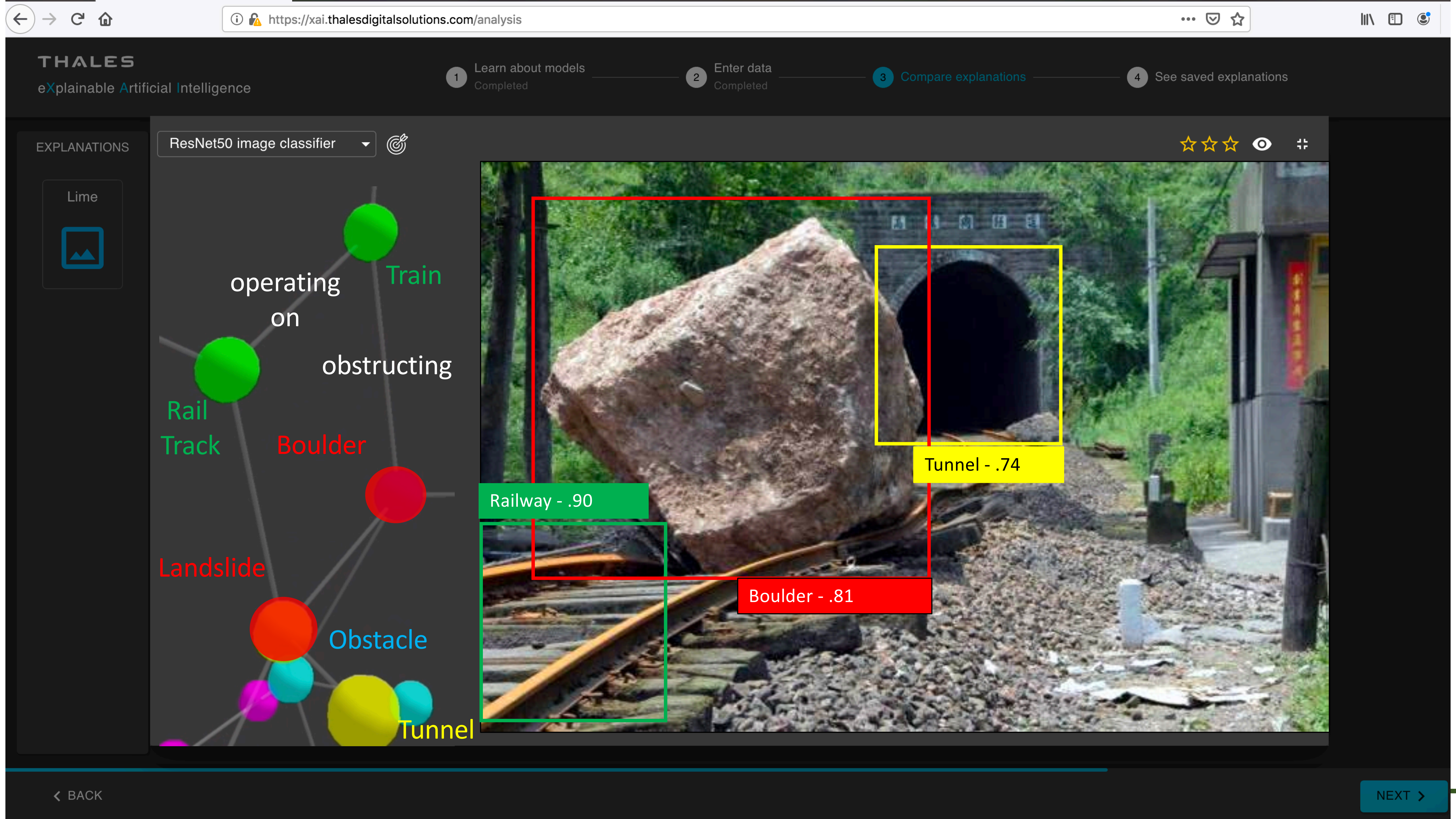


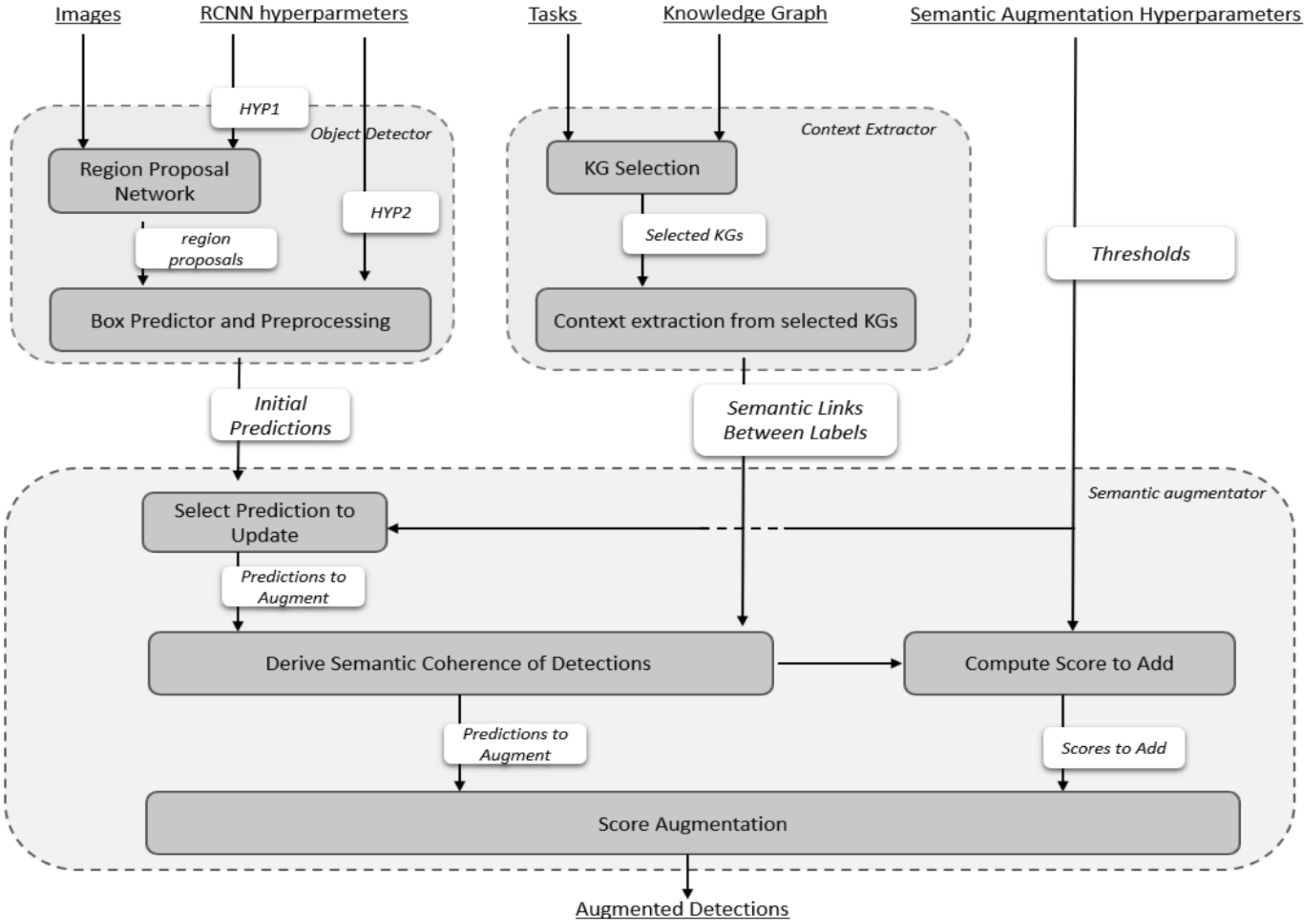
- **Hardware**: **High performance, scalable, generic** (to different FGPA family) **& portable** CNN dedicated **programmable** processor implemented on an FPGA for **real-time embedded inference**
- **Software**: Knowledge graph extension of object detection

Transitioning



This is an **Obstacle: Boulder** obstructing the train: XG142-R on **Rail_Track** from City: Cannes to City: Marseille at **Location: Tunnel VIX** due to **Landslide**


https://xai.thalesdigitalsolutions.com/analysis
THALES
eXplainable Artificial Intelligence
1 Learn about models
Completed
2 Enter data
Completed
3 Compare explanations
4 See saved explanations
EXPLANATIONS
Lime
ResNet50 image classifier
operating
on
Train
obstructing
Rail
Track
Boulder
Landslide
Obstacle
Tunnel
Tunnel - .74
Railway - .90
Boulder - .81
BACK
NEXT



Freddy Lécué, Jiaoyan Chen, Jeff Z. Pan, Huajun Chen: Augmenting Transfer Learning with Semantic Reasoning. IJCAI 2019: 1779-1785

Freddy Lécué, Tanguy Pommellet: Feeding Machine Learning with Knowledge Graphs for Explainable Object Detection. ISWC Satellites 2019: 277-280

Freddy Lécué, Baptiste Abeloos, Jonathan Anctil, Manuel Bergeron, Damien Dalla-Rosa, Simon Corbeil-Letourneau, Florian Martet, Tanguy Pommellet, Laura Salvan, Simon Veilleux, Maryam Ziaeefard: Thales XAI Platform: Adaptable Explanation of Machine Learning Systems - A Knowledge Graphs Perspective. ISWC Satellites 2019: 315-316

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

More on XAI

(Some) Tutorials, Workshops, Challenge

Tutorial:

- AAAI 2020 Tutorial On Explainable AI: From Theory to Motivation, Applications and Limitations (#2) - https://xaitutorial2019.github.io/ https://xaitutorial2020.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/

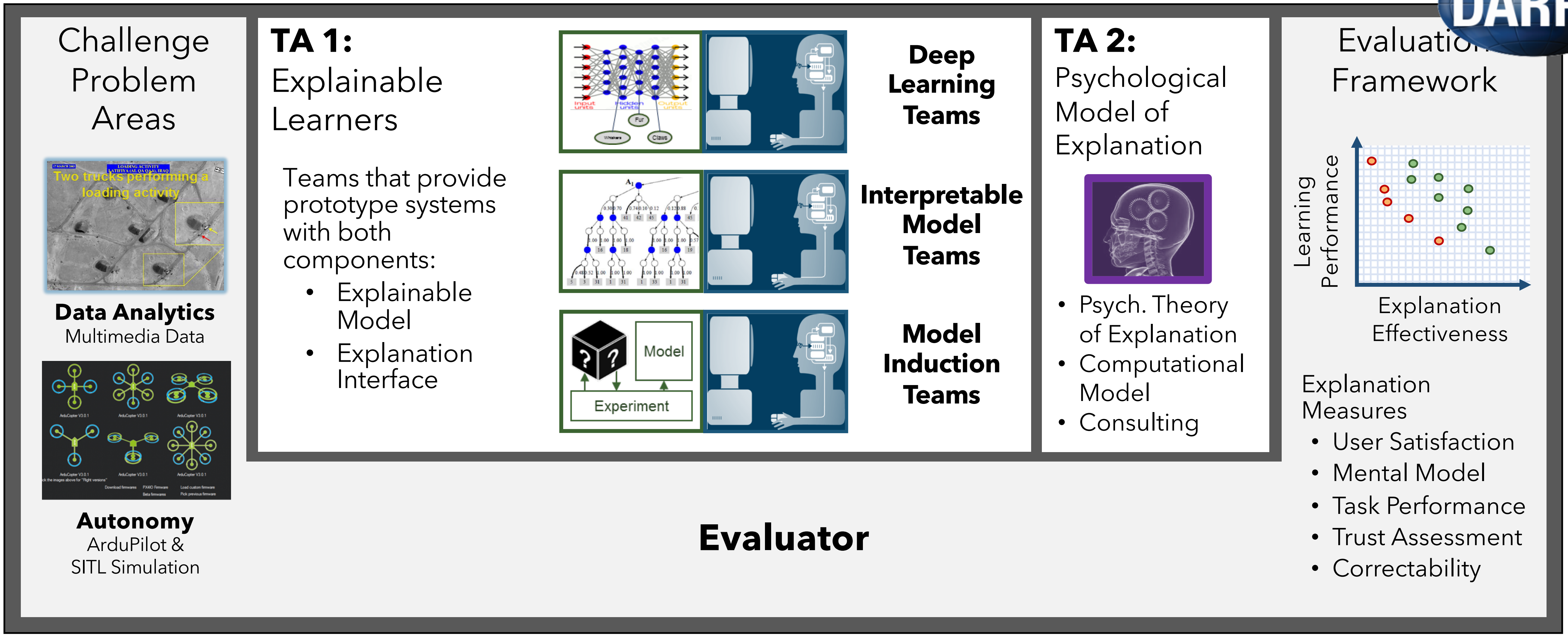
Challenge:

- 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



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 … ab … i … arning) Project 2019-2024
  - Research instit…
  - Indu… partners
  - A… ers

…nolo… w met… 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

Why do we Need Knowledge Graphs to Achieve XAI?

Because this is not an explanation from an intelligent system

This is even not interpretable, and then not actionable

Conclusion

- Explainable AI is motivated by **real-world applications in AI**
- Not a new problem – a reformulation of past research challenges in AI
- Knowledge graphs should be foundational for XAI
- But they are facing challenges related to their integration (data mapping)
- **Many industrial applications already – crucial for AI adoption in critical systems**

Open Research Questions for the Semantic Web / Knowledge Graph Community

- [Data] Machine learning experts do not buy the **data – knowledge mapping**
- [Explanation] There is ***no agreement*** on ***what an explanation is***
- [Explanation] There is ***not a formalism*** for ***explanations (neither model nor output)***
- [Model] *There is very limited work in **machine learning modules composability –** and **none from a semantics perspective***
- [Model] ***There is no work on describing and representing models***
- [Model] What are **disentangled representations** and how can its factors be quantified and detected?
- [Human-in-the-loop] There is ***no work*** that seriously addresses the problem of ***quantifying*** the grade of ***comprehensibility*** of an explanation for humans

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 1,800
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 Co
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 implemen
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 interest 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