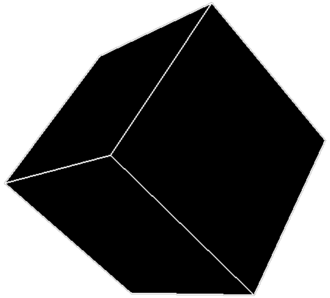
Explainable AI – The Story So Far



August 29th, 2019 - @Inha University

Freddy Lecue
Chief AI Scientist, CortAIx, Thales, Montreal – Canada
Inria, Sophia Antipolis - France

@freddylecue
https://tinyurl.com/freddylecue

Motivation

THALES

Business to Customer

THALES

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

THALES

Motivation (1)

Criminal Justice

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

Opinion

The New York Times

OP-ED CONTRIBUTOR

When a Computer Program Keeps You in Jail

By Rebecca Wexler

June 13, 2017

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

ACLU GET UPDATES / DONATE

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

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

How We Analyzed the COMPAS Recidivism Algorithm

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

May 23, 2016

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

THALES

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

THALES

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

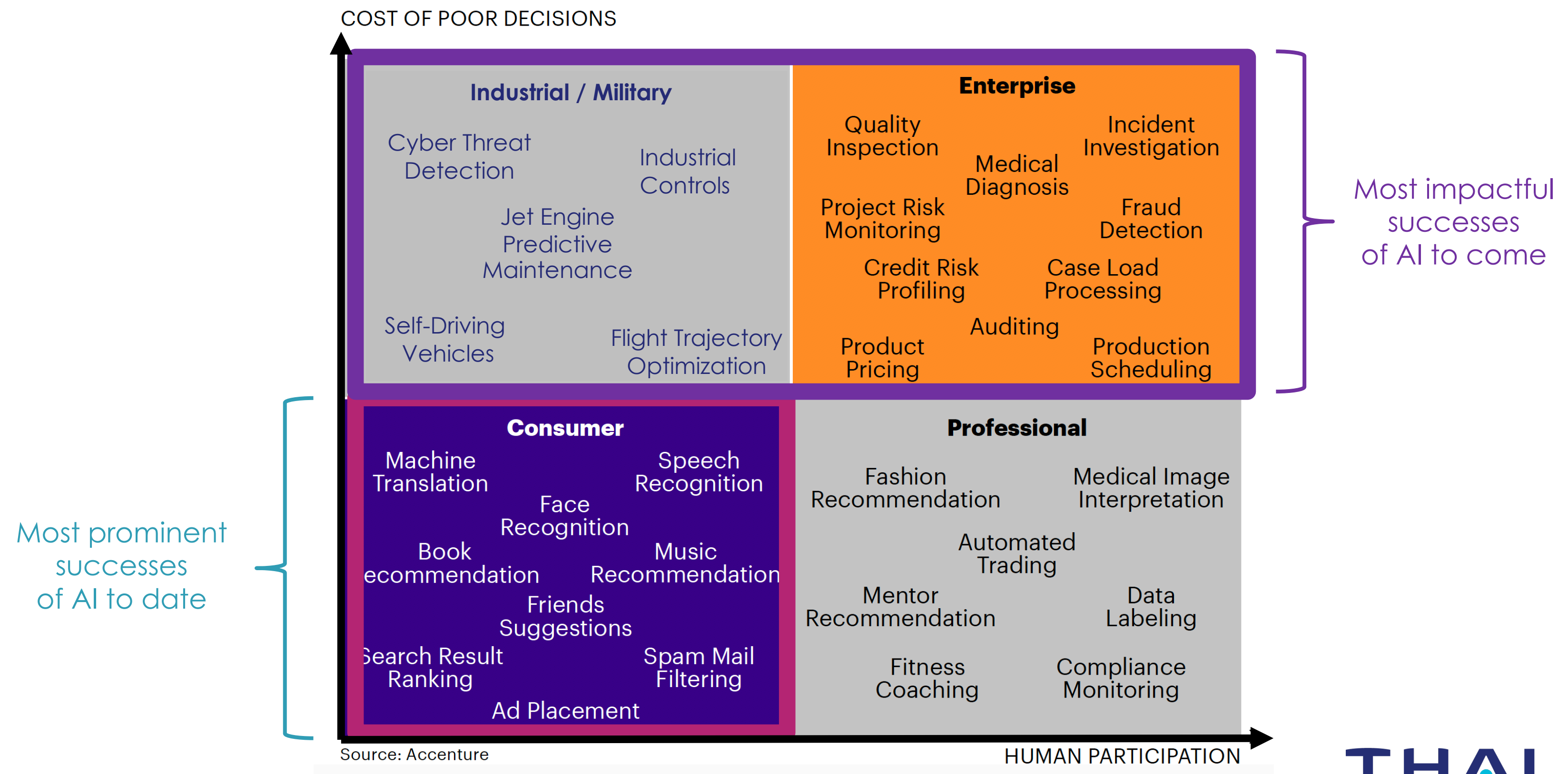
Marc Sturm
NewYork-Presbyterian Hospital
mas9161@nyp.org

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

Trustable AI and eXplainable AI: a Reality Need

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



Source: Accenture

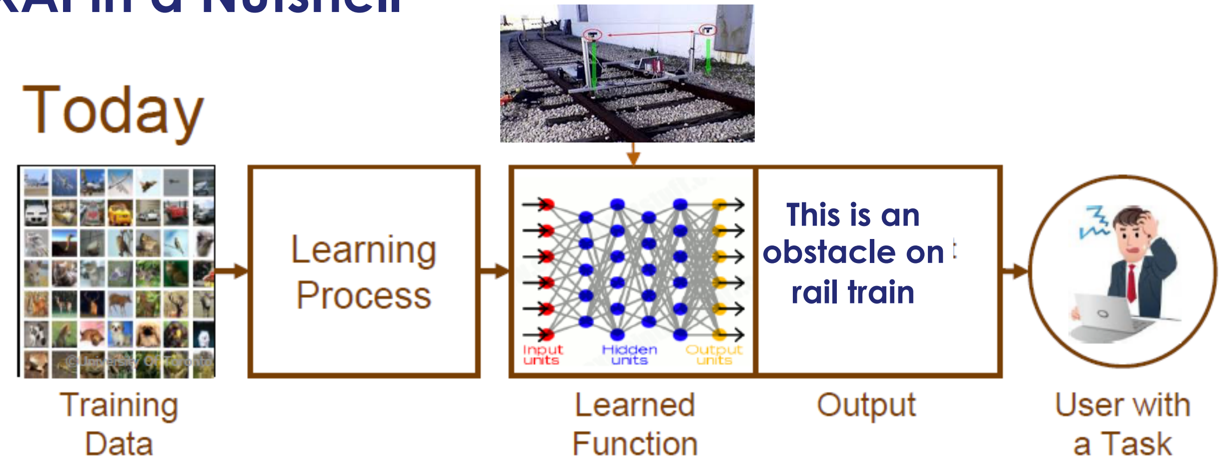
Source: Accenture Point of View. Understanding Machines: Explainable AI. Freddy Lecue, Dadong Wan

THALES

XAI in a Nutshell

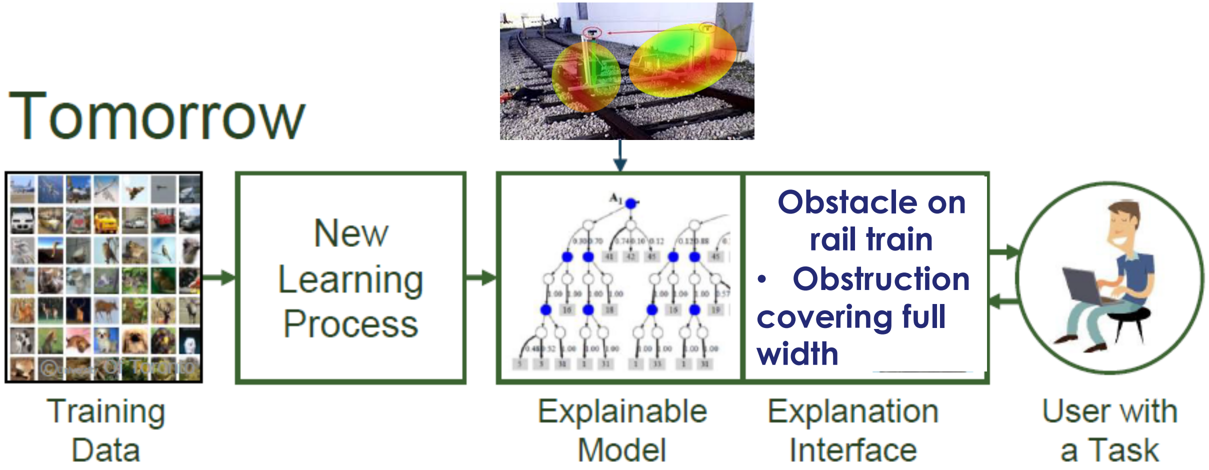
XAI in a Nutshell

Today



- 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?

Tomorrow



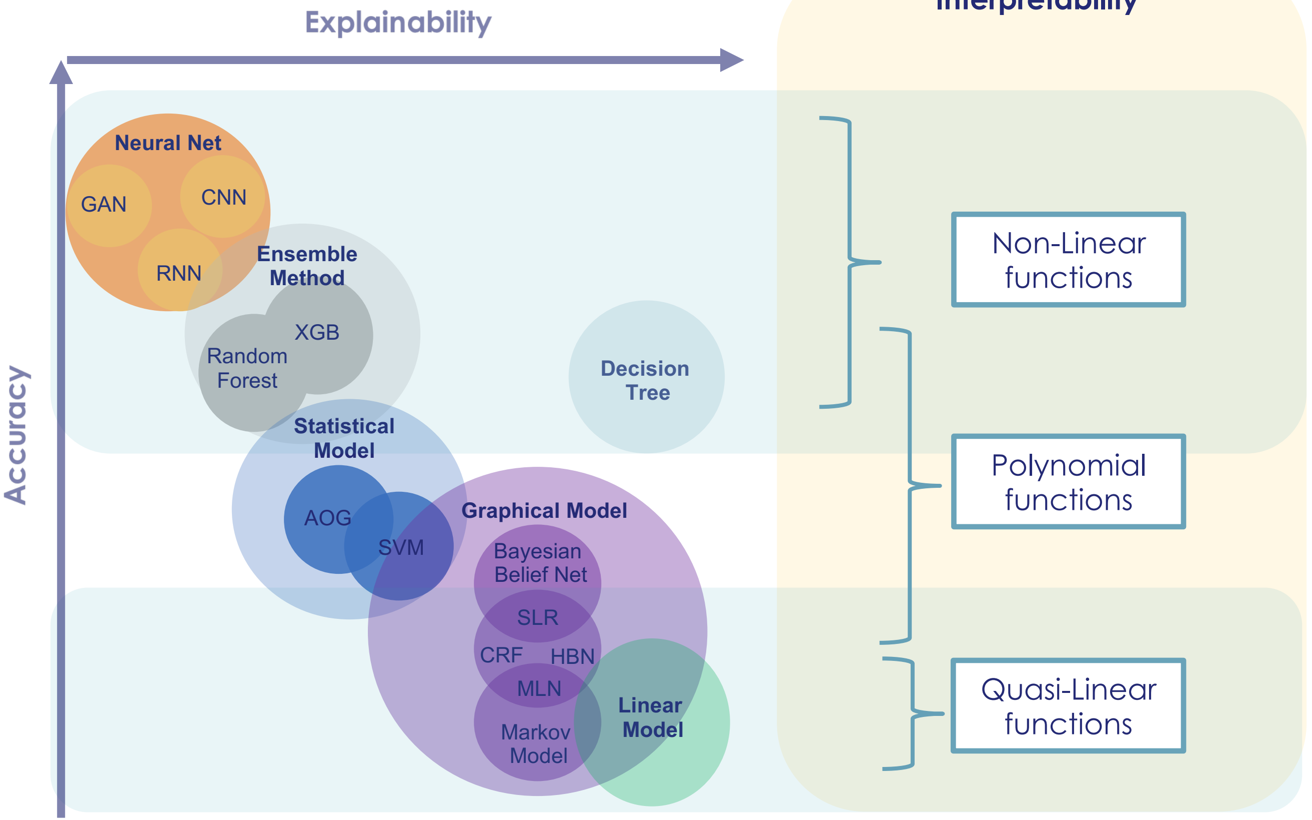
- I understand why
- I understand why not
- I know when you'll succeed
- I know when you'll fail
- I know when to trust you
- I know why you erred

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

How to Explain? Accuracy vs. Explainability

Learning

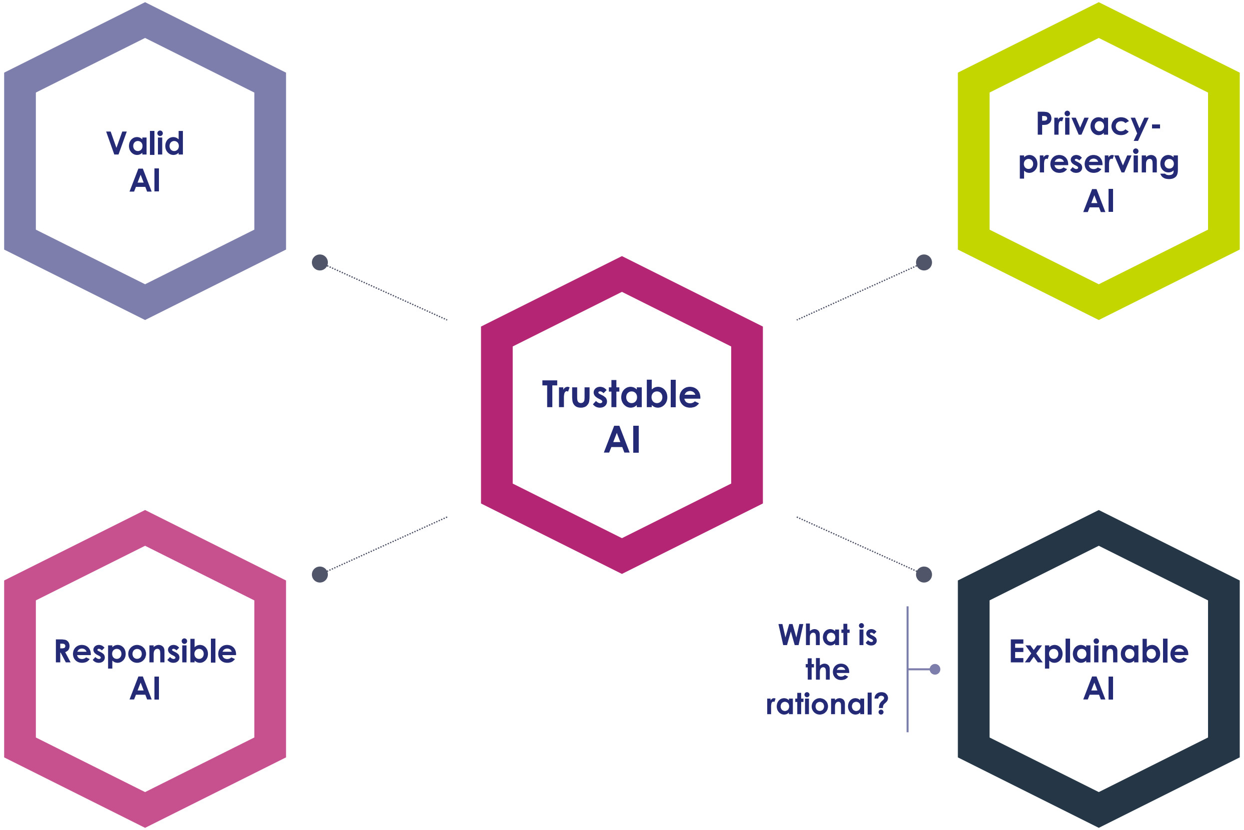
- Challenges:
  - Supervised
  - Unsupervised learning
- Approach:
  - Representation Learning
  - Stochastic selection
- Output:
  - **Correlation**
  - **No causation**



Trustable AI

THALES

AI Adoption: Requirements



- Human Interpretable AI
- Machine Interpretable AI

Definitions

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.

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.*

Models, Outputs of the Intelligent System

interpret | ɪnˈtəːprɪt |

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

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

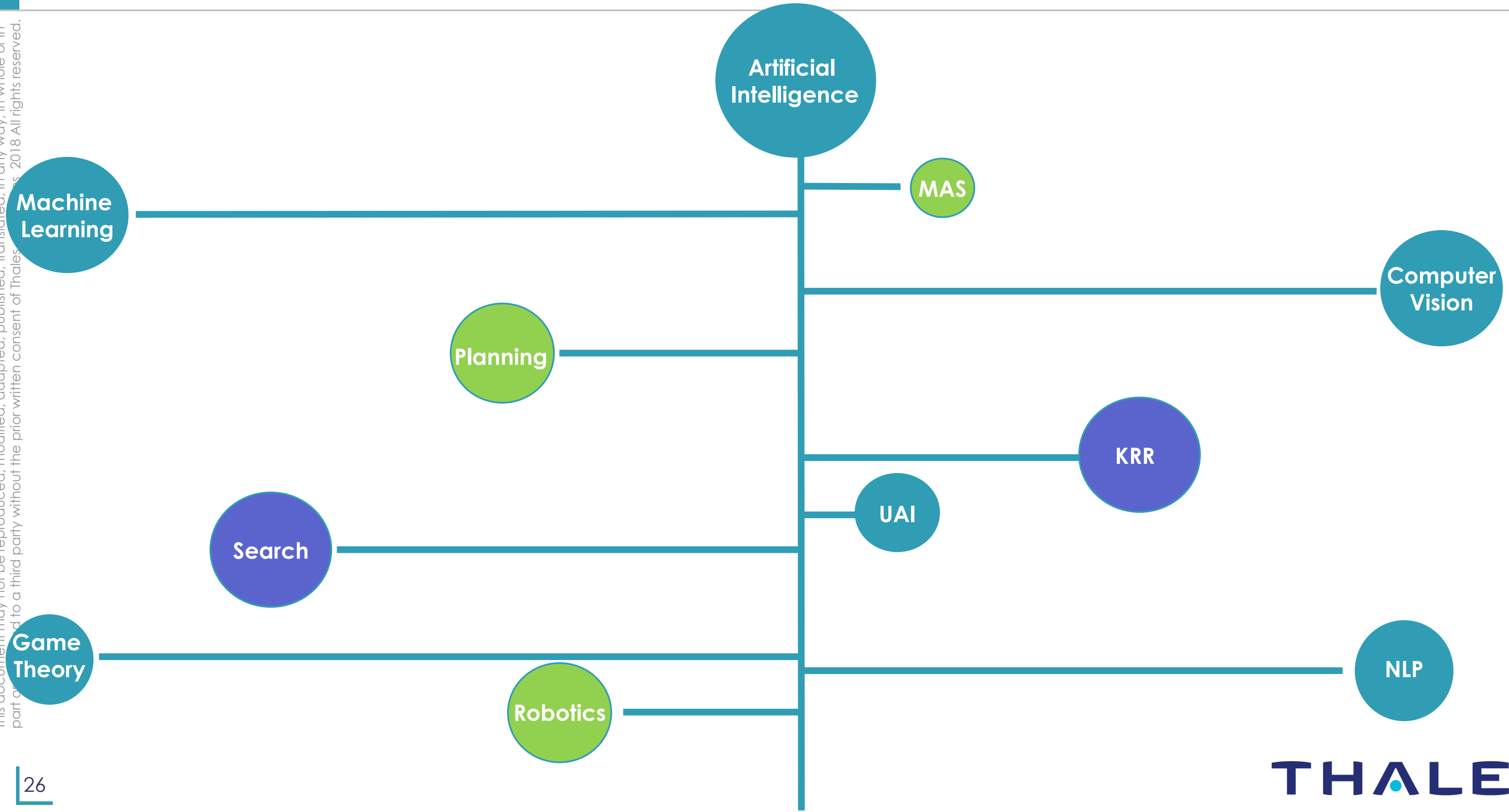
Models, Outputs of the Intelligent System

THALES

XAI in AI

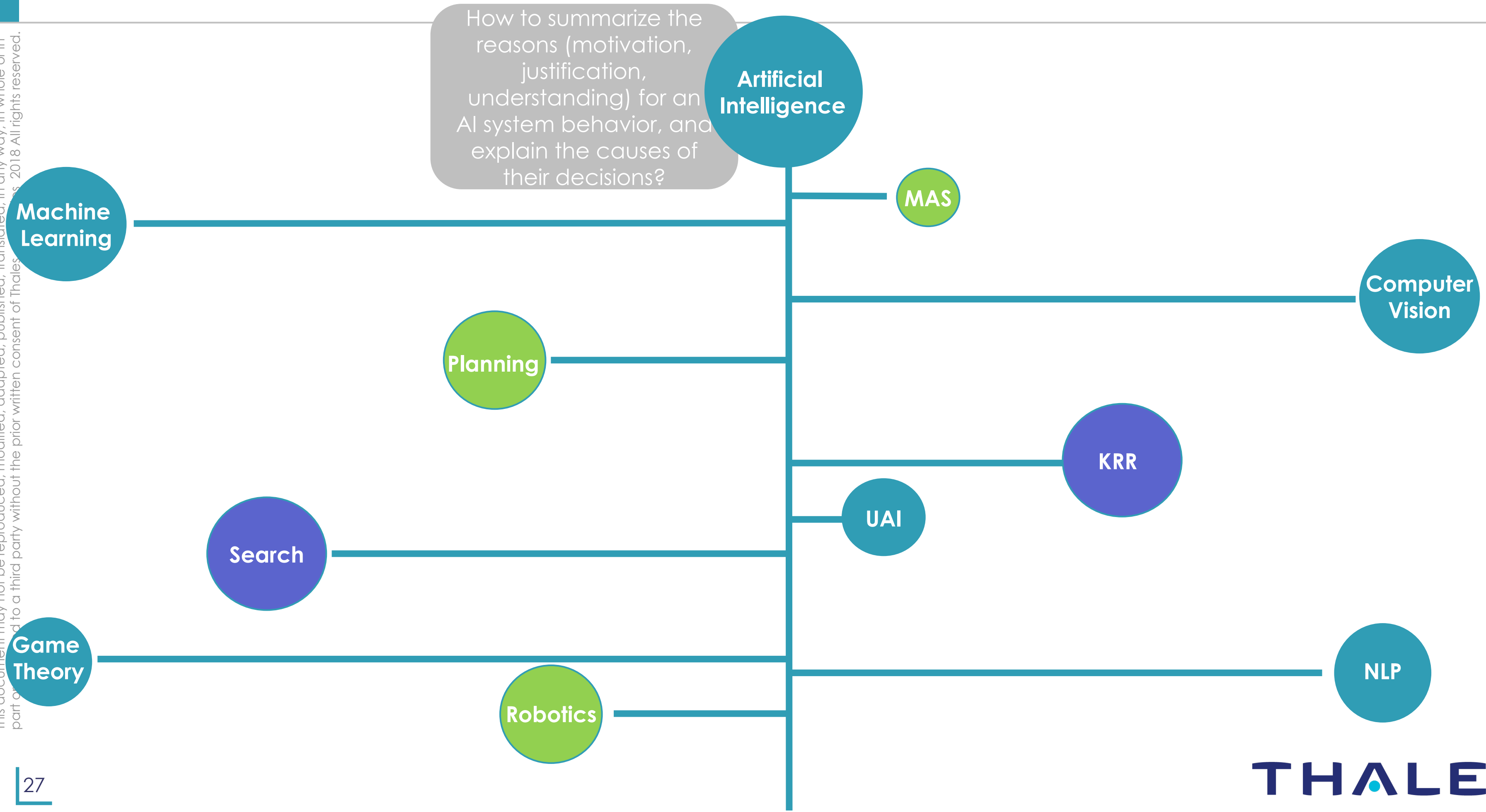
THALES

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



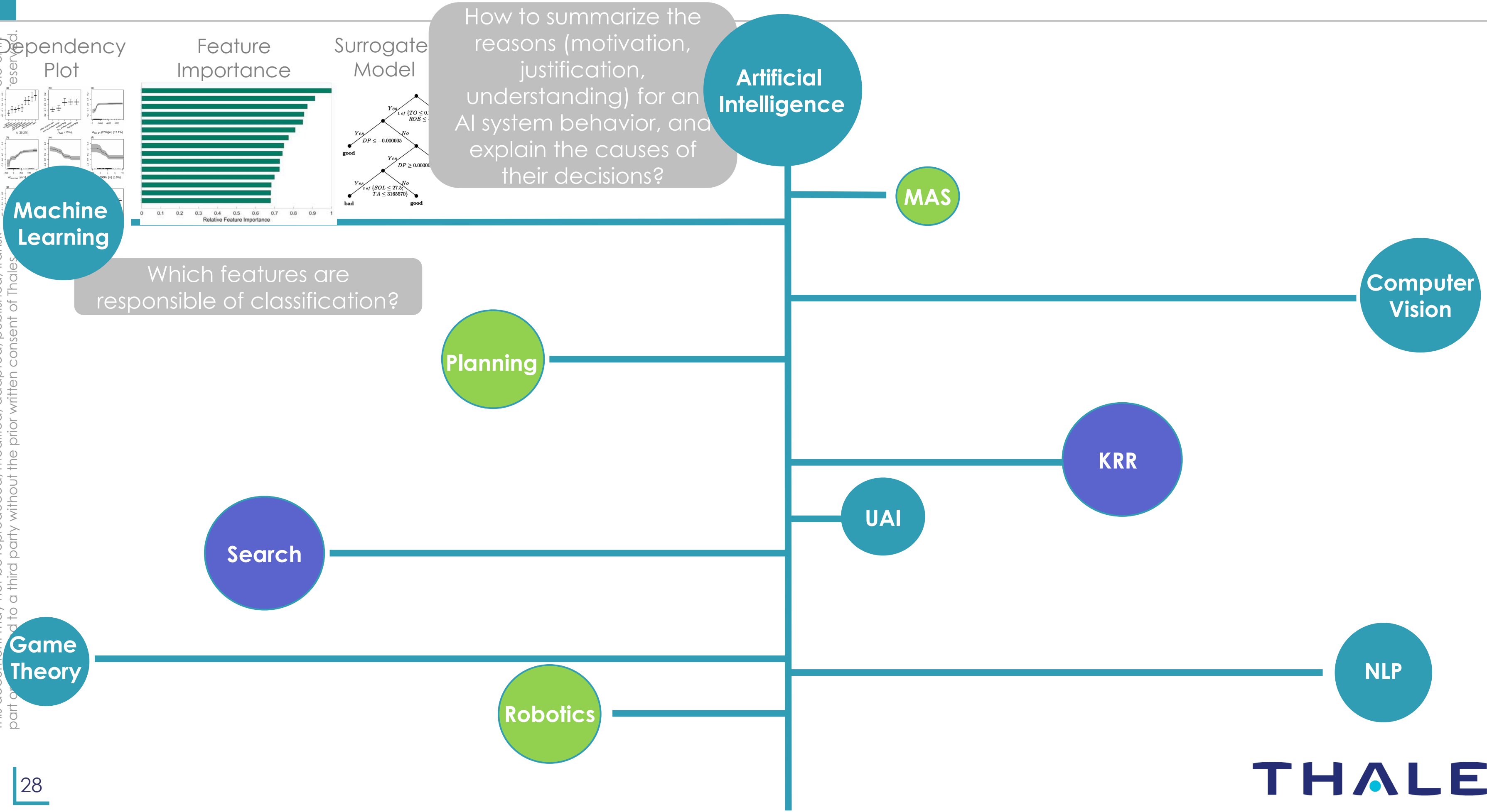
THALES

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



THALES

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



THALES

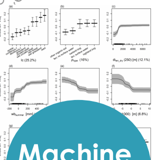
XAI: One Objective, Many 'AI's, Many Definitions, Many Approaches

Artificial Intelligence

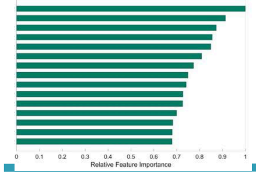
How to summarize the reasons (motivation, justification, understanding) for an AI system behavior, and explain the causes of their decisions?

Machine Learning

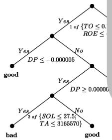
Dependency Plot



Feature Importance



Surrogate Model

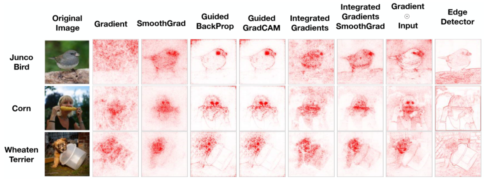


Which features are responsible of classification?

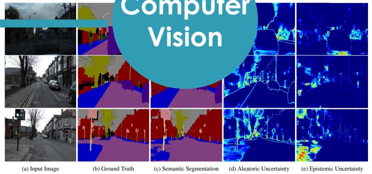
MAS

Computer Vision

Saliency Map



Which complex features are responsible of classification?



Uncertainty Map

Planning

KRR

UAI

Search

Game Theory

Robotics

NLP

THALES

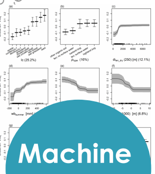
XAI: One Objective, Many 'AI's, Many Definitions, Many Approaches

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

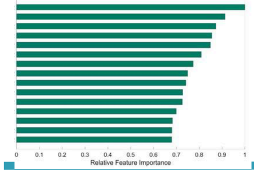
Artificial Intelligence

Machine Learning

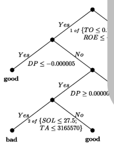
Dependency Plot



Feature Importance



Surrogate Model



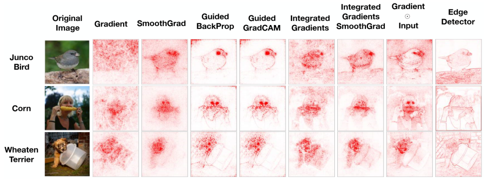
Which features are responsible of classification?

MAS

Strategy Summarization

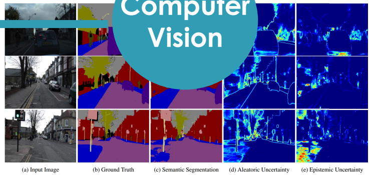
- Which agent strategy & plan ?
- Which player contributes most?
- Why such a conversational flow?

Saliency Map



Which complex features are responsible of classification?

Computer Vision



Uncertainty Map

Planning

KRR

UAI

Search

Game Theory

NLP

Robotics

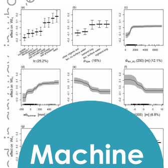
THALES

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

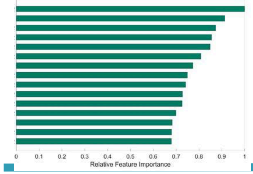
Artificial Intelligence

Machine Learning

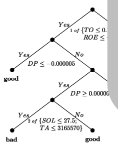
Dependency Plot



Feature Importance



Surrogate Model



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

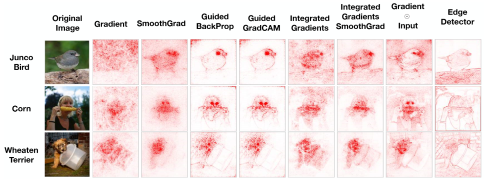
Which features are responsible of classification?

MAS

Strategy Summarization

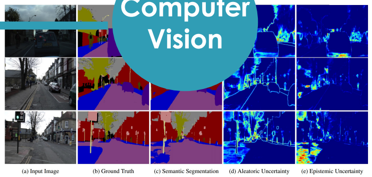
- Which agent strategy & plan ?
- Which player contributes most?
- Why such a conversational flow?

Saliency Map



Which complex features are responsible of classification?

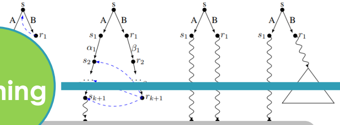
Computer Vision



Uncertainty Map

Planning

Plan Refinement



Which actions are responsible of a plan?

KRR

UAI

Search

Game Theory

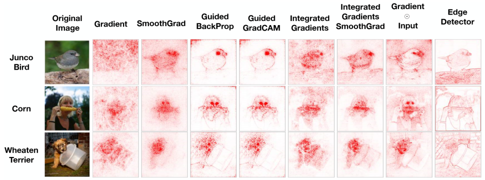
Robotics

NLP

THALES

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

Saliency Map



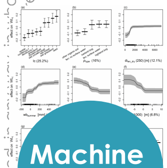
Which complex features are responsible of classification?

Strategy Summarization

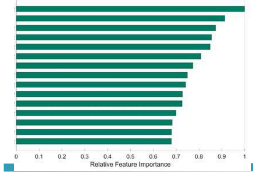
Artificial Intelligence

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

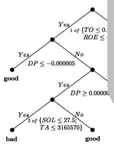
Dependency Plot



Feature Importance



Surrogate Model



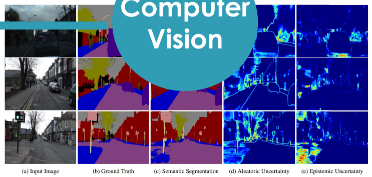
Machine Learning

Which features are responsible of classification?

MAS

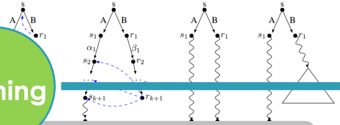
- Which agent strategy & plan ?
- Which player contributes most?
- Why such a conversational flow?

Computer Vision



Uncertainty Map

Plan Refinement

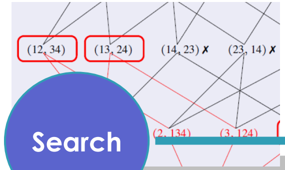


Planning

Which actions are responsible of a plan?

KRR

Conflicts Resolution



UAI

Search

Which constraints can be relaxed?

Game Theory

NLP

Robotics

THALES

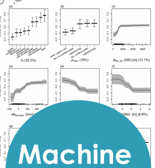
XAI: One Objective, Many 'AI's, Many Definitions, Many Approaches

Artificial Intelligence

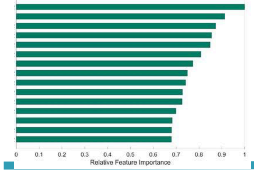
How to summarize the reasons (motivation, justification, understanding) for an AI system behavior, and explain the causes of their decisions?

Machine Learning

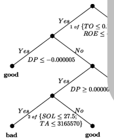
Dependency Plot



Feature Importance



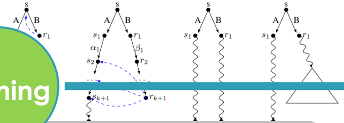
Surrogate Model



Which features are responsible of classification?

Planning

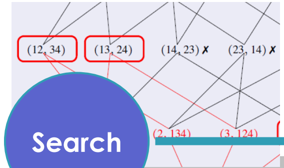
Plan Refinement



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

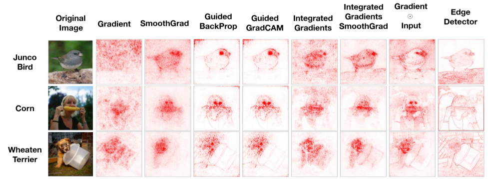
MAS

Strategy Summarization

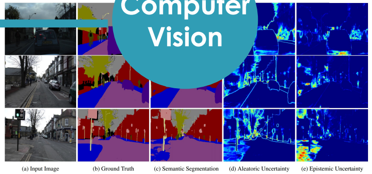
- Which agent strategy & plan ?
- Which player contributes most?
- Why such a conversational flow?

Computer Vision

Saliency Map



Which complex features are responsible of classification?



Uncertainty Map

KRR

UAI

NLP

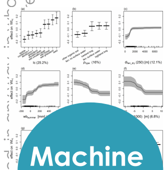
THALES

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

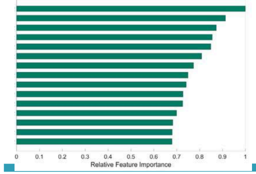
How to summarize the reasons (motivation, justification, understanding) for an AI system behavior, and explain the causes of their decisions?

Artificial Intelligence

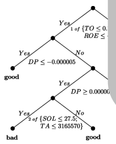
Dependency Plot



Feature Importance



Surrogate Model



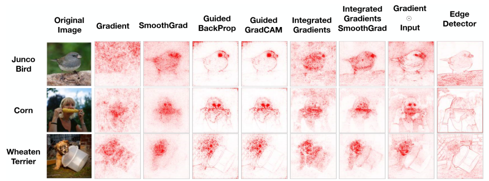
Machine Learning

Which features are responsible of classification?

Strategy Summarization

MAS

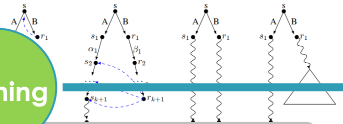
Saliency Map



Which complex features are responsible of classification?

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

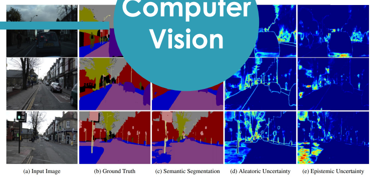
Plan Refinement



Planning

Which actions are responsible of a plan?

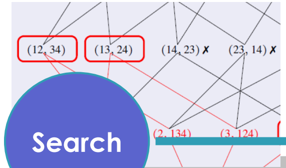
Computer Vision



Uncertainty Map

KRR

Conflicts Resolution



UAI

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?

NLP

Narrative-based

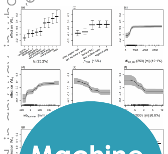
THALES

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

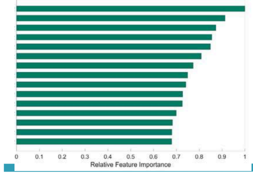
Artificial Intelligence

Machine Learning

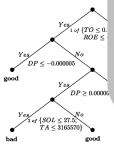
Dependency Plot



Feature Importance



Surrogate Model



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?

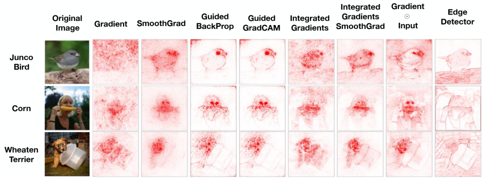
MAS

Strategy Summarization

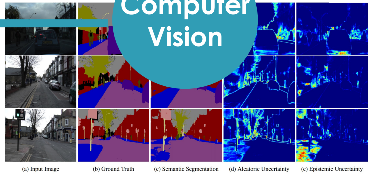
- Which agent strategy & plan ?
- Which player contributes most?
- Why such a conversational flow?

Computer Vision

Saliency Map



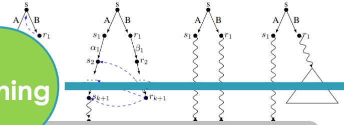
Which complex features are responsible of classification?



Uncertainty Map

Planning

Plan Refinement



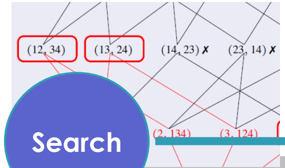
Which actions are responsible of a plan?

KRR

UAI

Search

Conflicts Resolution



Which constraints can be relaxed?

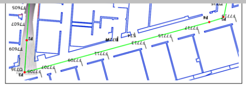
Game Theory

Which combination of features is optimal?

Shapely Values

Robotics

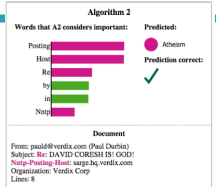
Which decisions, combination of multimodal decisions lead to an action?



Narrative-based

NLP

Machine Learning based



Which entity is responsible for classification?

THALES

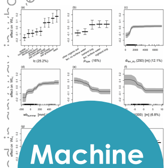
XAI: One Objective, Many 'AI's, Many Definitions, Many Approaches

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

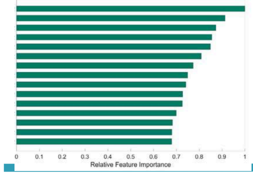
Artificial Intelligence

Machine Learning

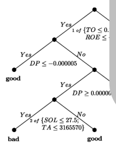
Dependency Plot



Feature Importance



Surrogate Model



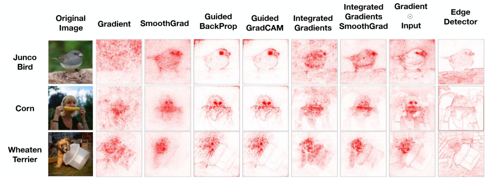
Which features are responsible of classification?

MAS

Strategy Summarization

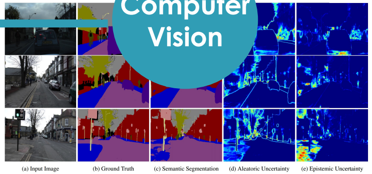
- Which agent strategy & plan ?
- Which player contributes most?
- Why such a conversational flow?

Saliency Map



Which complex features are responsible of classification?

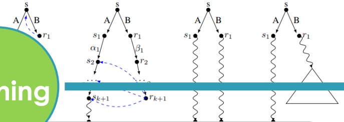
Computer Vision



Uncertainty Map

Planning

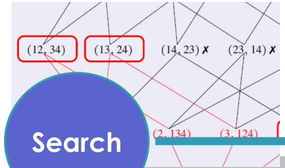
Plan Refinement



Which actions are responsible of a plan?

Search

Conflicts Resolution

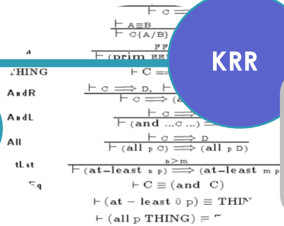


Which constraints can be relaxed?

KRR

UAI

Diagnosis



Abduction

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

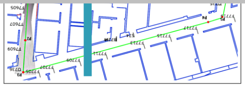
Game Theory

Which combination of features is optimal?

Shapely Values

Robotics

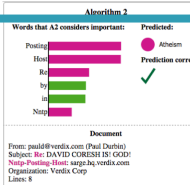
Which decisions, combination of multimodal decisions lead to an action?



Narrative-based

NLP

Machine Learning based



Which entity is responsible for classification?

THALES

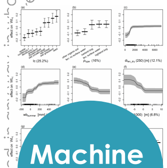
XAI: One Objective, Many 'AI's, Many Definitions, Many Approaches

Artificial Intelligence

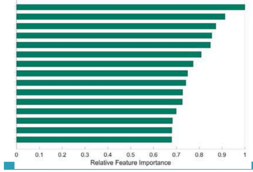
How to summarize the reasons (motivation, justification, understanding) for an AI system behavior, and explain the causes of their decisions?

Machine Learning

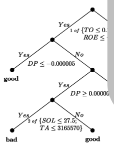
Dependency Plot



Feature Importance



Surrogate Model



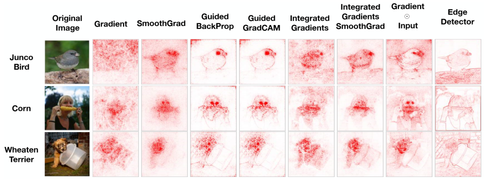
Which features are responsible of classification?

MAS

Strategy Summarization

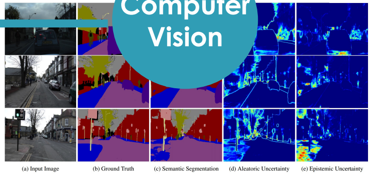
- Which agent strategy & plan ?
- Which player contributes most?
- Why such a conversational flow?

Saliency Map



Which complex features are responsible of classification?

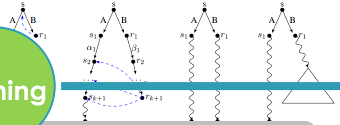
Computer Vision



Uncertainty Map

Planning

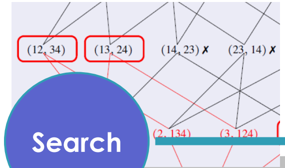
Plan Refinement



Which actions are responsible of a plan?

Search

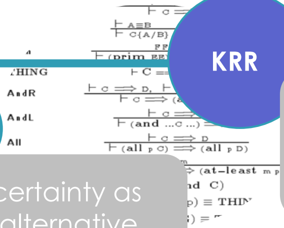
Conflicts Resolution



Which constraints can be relaxed?

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

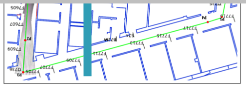
Game Theory

Which combination of features is optimal?

Shapely Values

Robotics

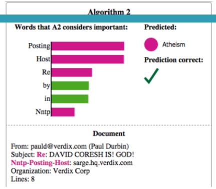
Which decisions, combination of multimodal decisions lead to an action?



Narrative-based

NLP

Machine Learning based



Which entity is responsible for classification?

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

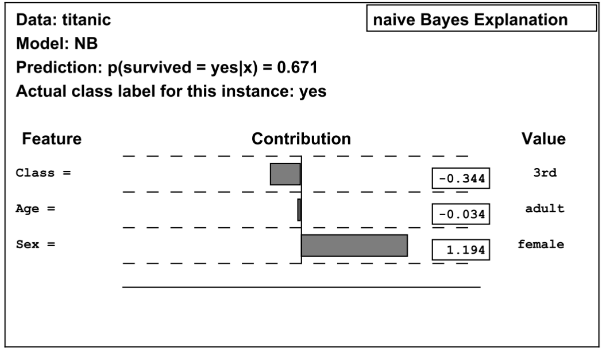
THALES

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.

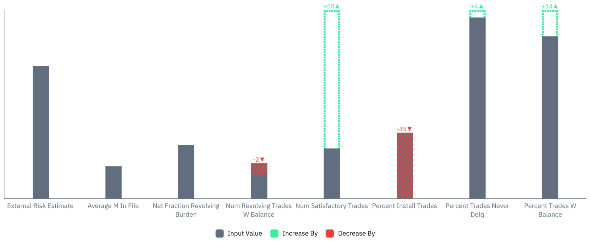
THALES

Overview of explanation in different AI fields (1)

Machine Learning (except Artificial Neural Network)

Interpretable Models:

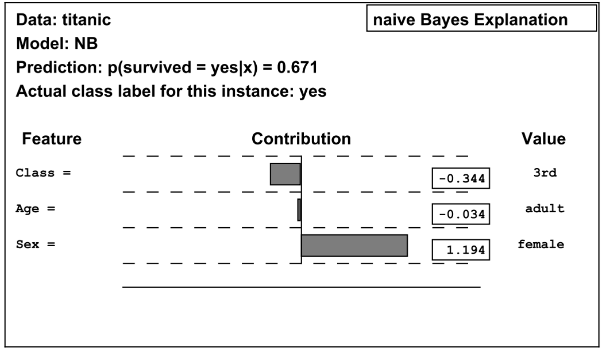
- Linear regression,
- Logistic regression,
- Decision Tree,
- GLMs,
- GAMs
- KNNs



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)



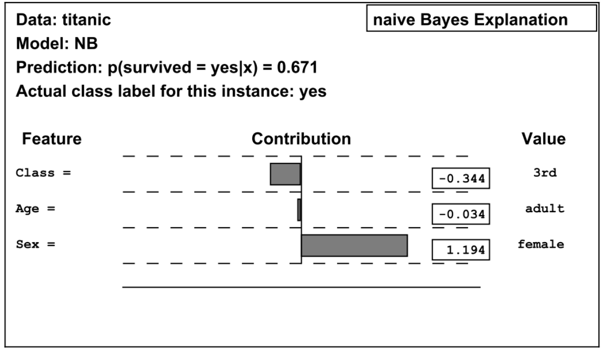
Naive Bayes model

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

Overview of explanation in different AI fields (1)

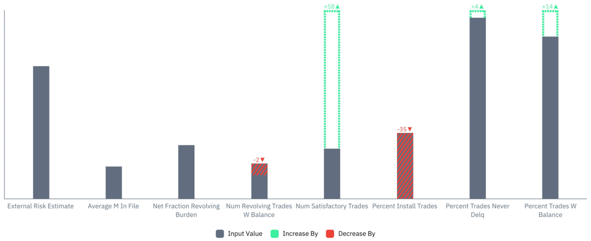
Machine Learning (except Artificial Neural Network)

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



Naive Bayes model

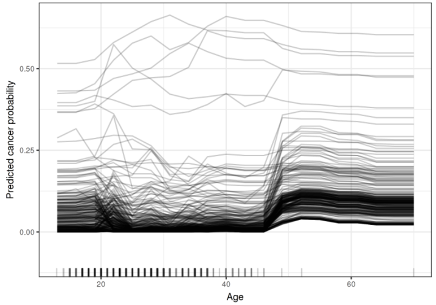
Igor Kononenko. Machine learning for medical diagnosis: history, state of the art and perspective. Artificial Intelligence in Medicine, 23:89–109, 2001.

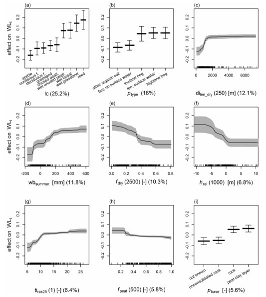


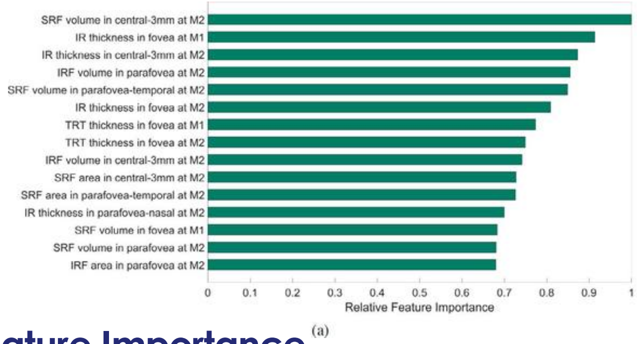
Counterfactual What-if

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)





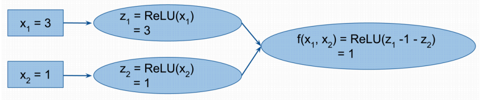


Feature Importance
Partial Dependence Plot
Individual Conditional Expectation
Sensitivity Analysis

THALES

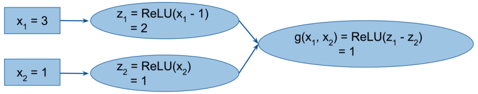
Overview of explanation in different AI fields (2)

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)

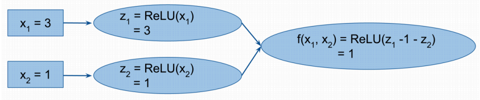
Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In ICML, pp. 3319–3328, 2017.

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

THALES

Overview of explanation in different AI fields (2)

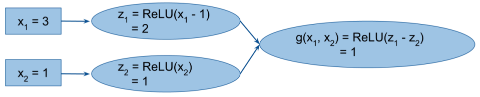
Machine Learning (only Artificial Neural Network)



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 $g(x_1, x_2)$

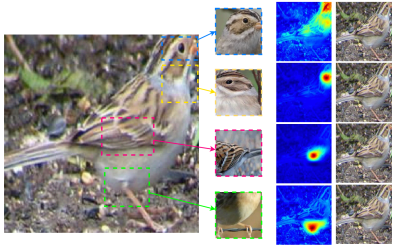
Attributions at $x_1 = 3, x_2 = 1$

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

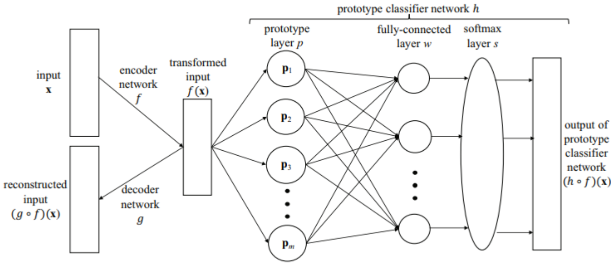
Attribution for Deep Network (Integrated gradient-based)

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



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)



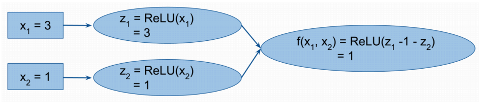
Auto-encoder / Prototype

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

THALES

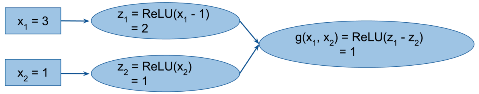
Overview of explanation in different AI fields (2)

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	$x_1 = 1.5,\ x_2 = -0.5$



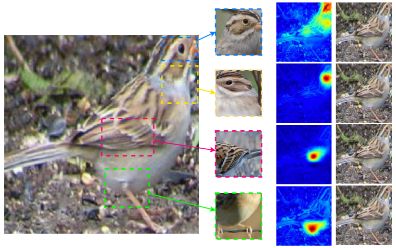
Network $g(x_1, x_2)$
Attributions at $x_1 = 3, x_2 = 1$

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

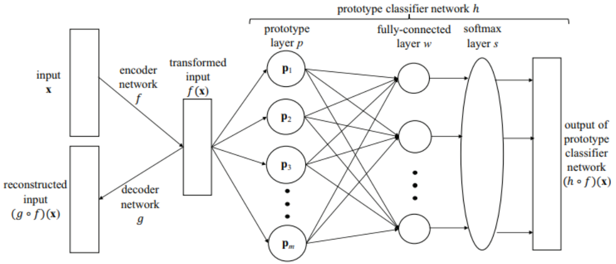
Attribution for Deep Network (Integrated gradient-based)

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

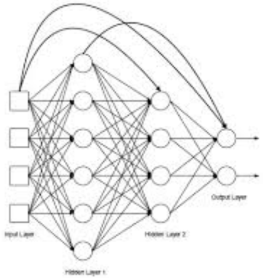


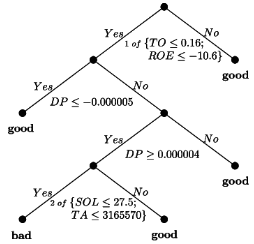
Chaofan Chen, Oscar Li, Alina Barnett, Jonathan Su, Cynthia Rudin: This looks like that: deep learning for interpretable image recognition. CoRR abs/1806.10574 (2018)



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



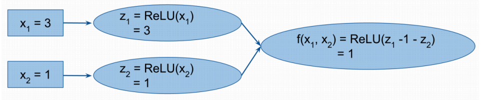


Surogate Model

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

Overview of explanation in different AI fields (2)

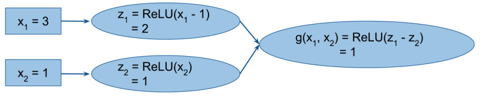
Machine Learning (only Artificial Neural Network)



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 $g(x_1, x_2)$

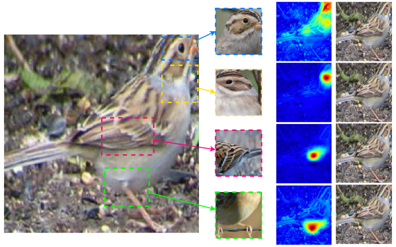
Attributions at $x_1 = 3, x_2 = 1$

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

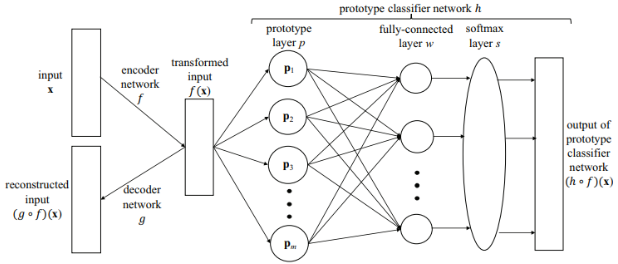
Attribution for Deep Network (Integrated gradient-based)

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

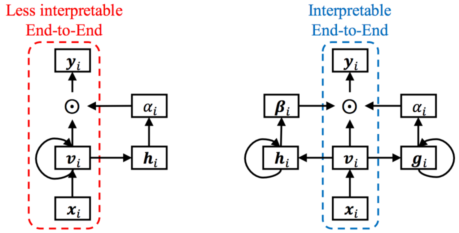


Chaofan Chen, Oscar Li, Alina Barnett, Jonathan Su, Cynthia Rudin: This looks like that: deep learning for interpretable image recognition. CoRR abs/1806.10574 (2018)



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

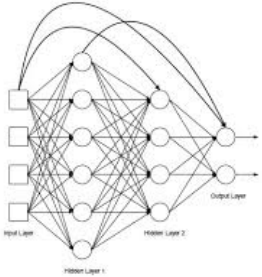


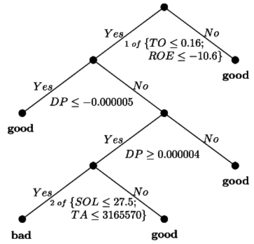
(a) Standard attention model (b) RETAIN model

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





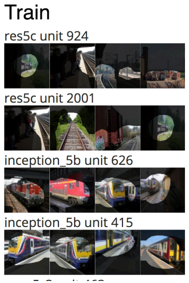
Surogate Model

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

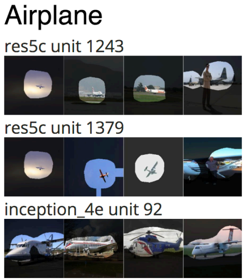
THALES

Overview of explanation in different AI fields (3)

Computer Vision



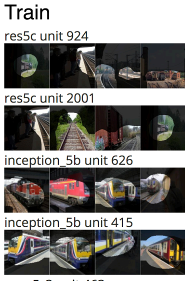
Interpretable Units



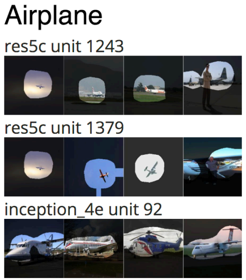
David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, Antonio Torralba: Network Dissection: Quantifying Interpretability of Deep Visual Representations. CVPR 2017: 3319-3327

Overview of explanation in different AI fields (3)

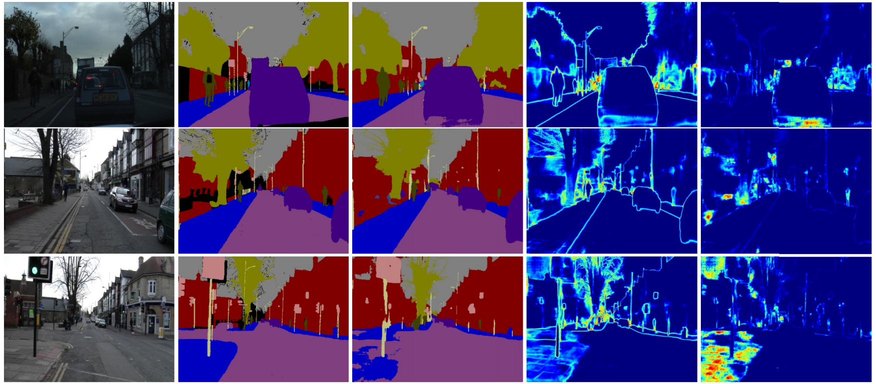
Computer Vision



Interpretable Units



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



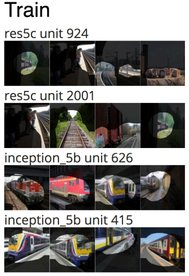
(a) Input Image (b) Ground Truth (c) Semantic Segmentation (d) Aleatoric Uncertainty (e) Epistemic Uncertainty

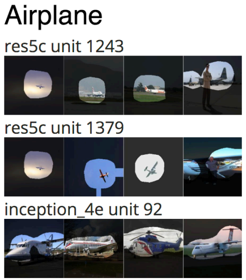
Uncertainty Map

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)

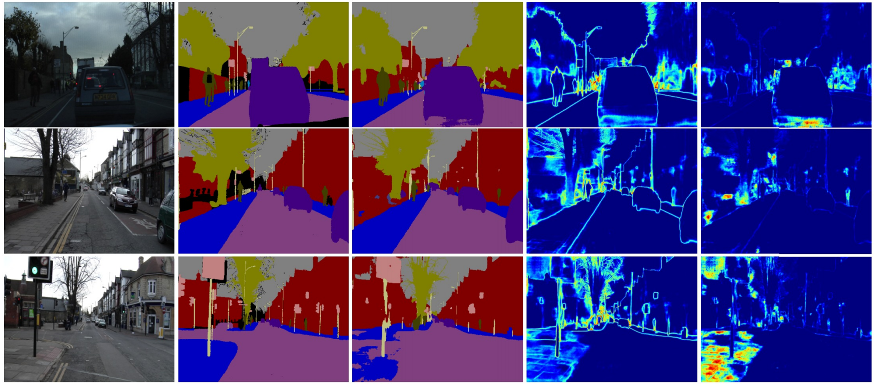
Computer Vision





Interpretable Units

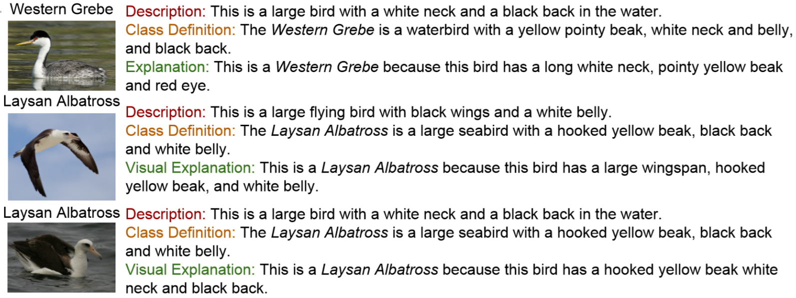
David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, Antonio Torralba: Network Dissection: Quantifying Interpretability of Deep Visual Representations. CVPR 2017: 3319-3327



(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



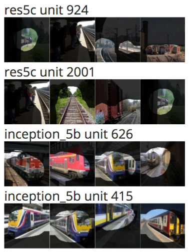
Visual Explanation

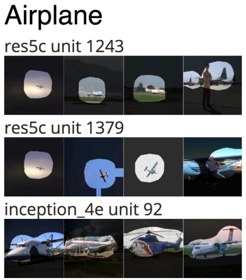
Lisa Anne Hendricks, Zeynep Akata, Marcus Rohrbach, Jeff Donahue, Bernt Schiele, Trevor Darrell: Generating Visual Explanations. ECCV (4) 2016: 3-19

THALES

Overview of explanation in different AI fields (3)

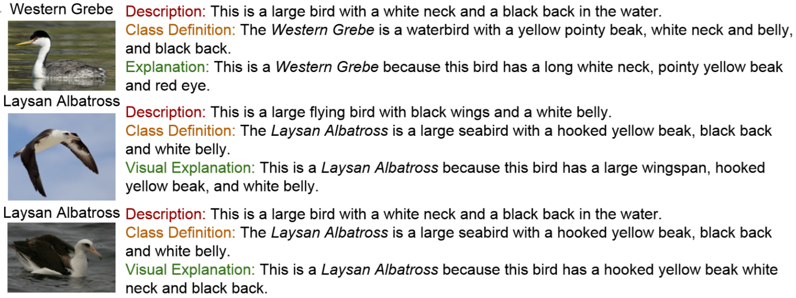
Computer Vision





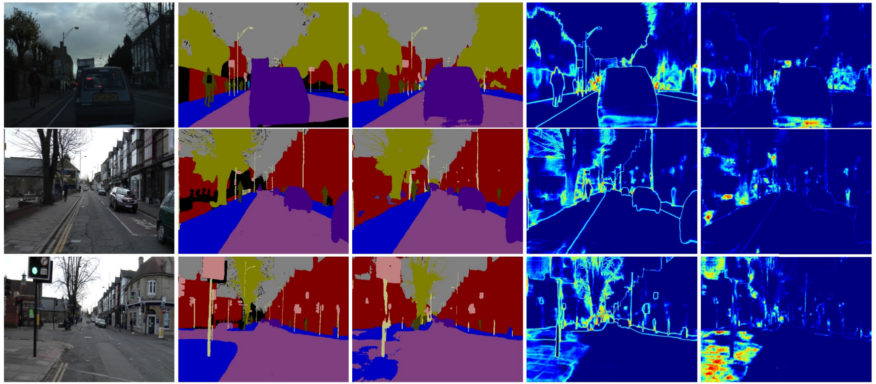
Interpretable Units

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



Visual Explanation

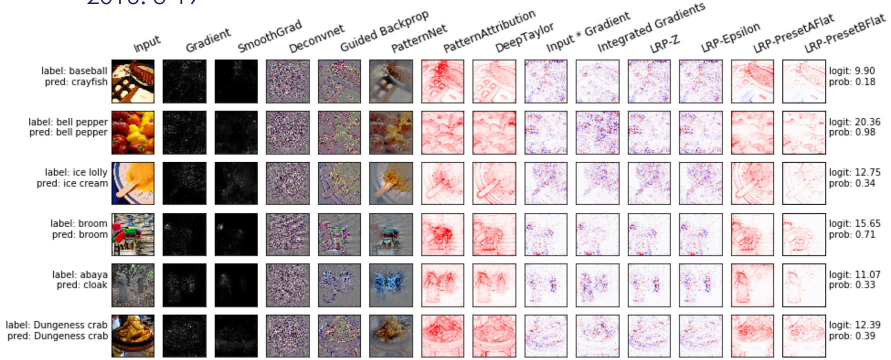
Lisa Anne Hendricks, Zeynep Akata, Marcus Rohrbach, Jeff Donahue, Bernt Schiele, Trevor Darrell: Generating Visual Explanations. ECCV (4) 2016: 3-19



(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

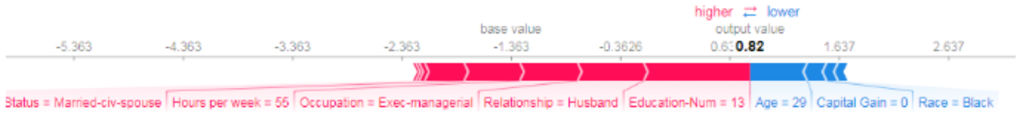


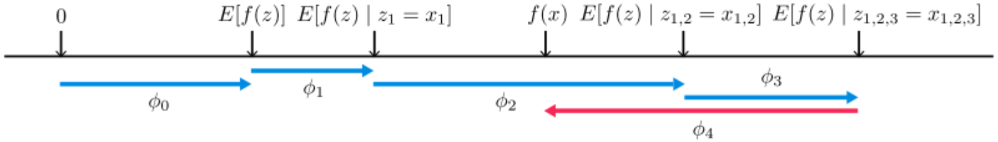
Saliency Map

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



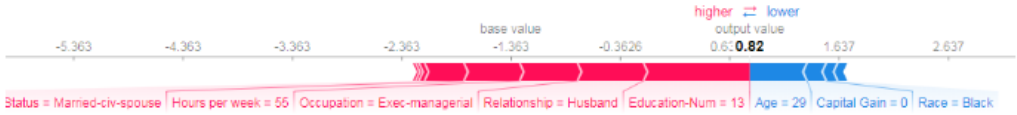


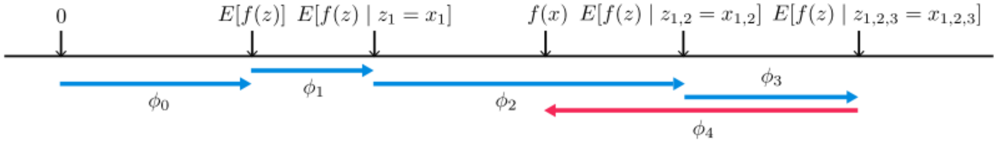
Shapley Additive Explanation

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)

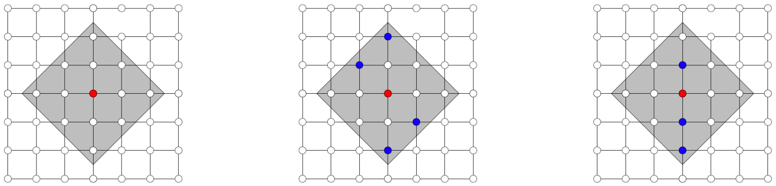
Game Theory





Shapley Additive Explanation

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

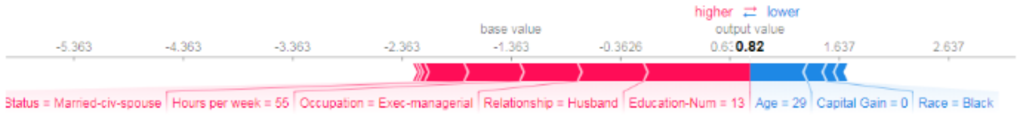


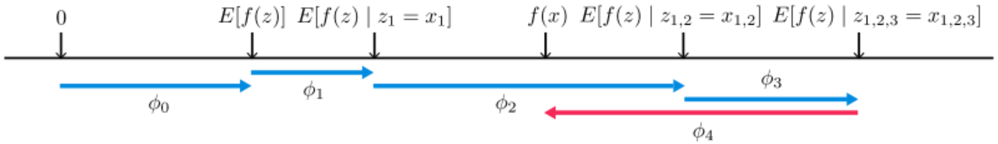
L-Shapley and C-Shapley (with graph structure)

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)

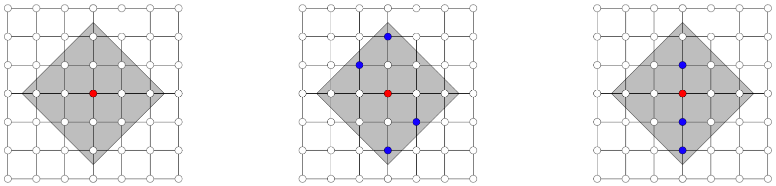
Game Theory





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.

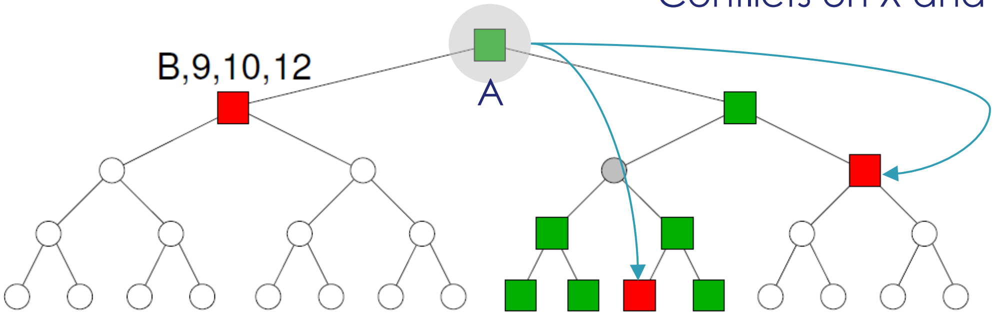
Anupam Datta, Shayak Sen, and Yair Zick. Algorithmic transparency via quantitative input influence: Theory and experiments with learning systems. In Security and Privacy (SP), 2016 IEEE Symposium on, pp. 598–617. IEEE, 2016.

THALES

Overview of explanation in different AI fields (5)

Search and Constraint Satisfaction

If A+1 then NEW
Conflicts on X and Y



Conflicts resolution

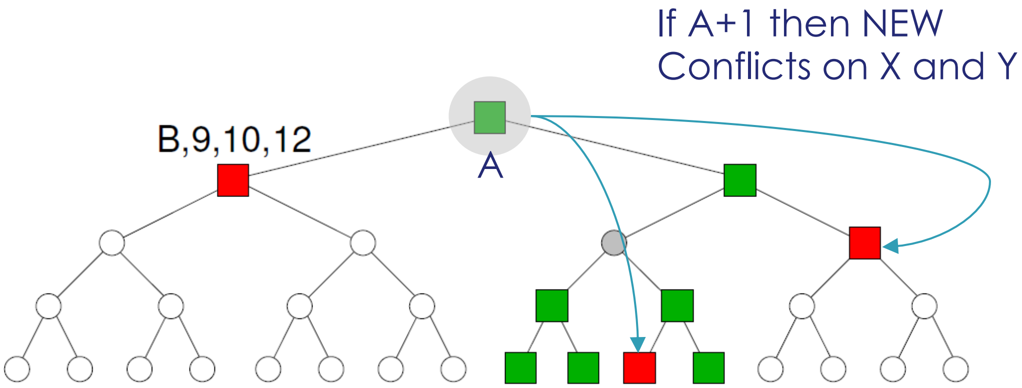
Barry O'Sullivan, Alexandre Papadopoulos, Boi Faltings, Pearl Pu: Representative Explanations for Over-Constrained Problems. AAAI 2007: 323-328

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

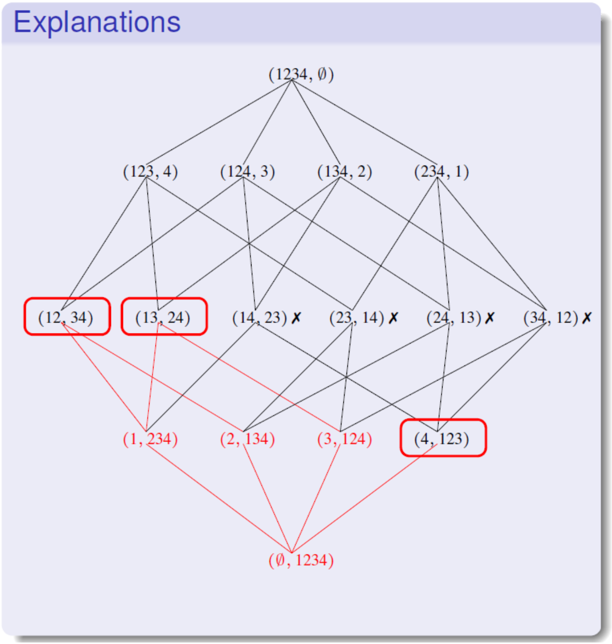


Conflicts resolution

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

THALES

Overview of explanation in different AI fields (6)

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

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

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)

Knowledge Representation and Reasoning

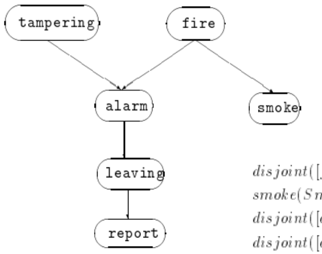
Ref	$\vdash C \Longrightarrow C$
Trans	$\dfrac{\vdash C \Longrightarrow D,\ \vdash D \Longrightarrow E}{\vdash C \Longrightarrow E}$
Eq	$\dfrac{\vdash A \equiv B \quad \vdash C \Longrightarrow D}{\vdash C\{A/B\} \Longrightarrow D\{A/B\}}$
Prim	$\dfrac{FF \subset EE}{\vdash (\text{prim } EE) \Longrightarrow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

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

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

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

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



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

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

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)

Knowledge Representation and Reasoning

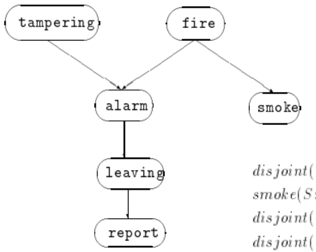
Ref	$\vdash C \Longrightarrow C$
Trans	$\dfrac{\vdash C \Longrightarrow D,\ \vdash D \Longrightarrow E}{\vdash C \Longrightarrow E}$
Eq	$\dfrac{\vdash A \equiv B \qquad \vdash C \Longrightarrow D}{\vdash C\{A/B\} \Longrightarrow D\{A/B\}}$
Prim	$\dfrac{FF \subset EE}{\vdash (\mathbf{prim}\ EE) \Longrightarrow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

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

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

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

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

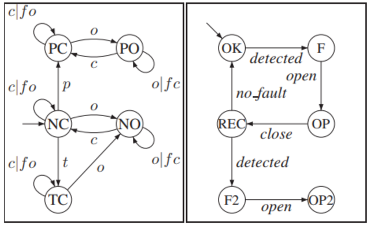


$$
\begin{aligned}
P(alarm|fire \wedge \neg tampering) &= 0.99 \\
P(alarm|\neg fire \wedge tampering) &= 0.85 \\
P(alarm|\neg fire \wedge \neg tampering) &= 0.0001 \\
P(leaving|alarm) &= 0.88 \\
P(leaving|\neg alarm) &= 0.001 \\
P(\mathrm{report}|leaving) &= 0.75 \\
P(\mathrm{report}|\neg leaving) &=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

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

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

THALES

Overview of explanation in different AI fields (7)

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)

THALES

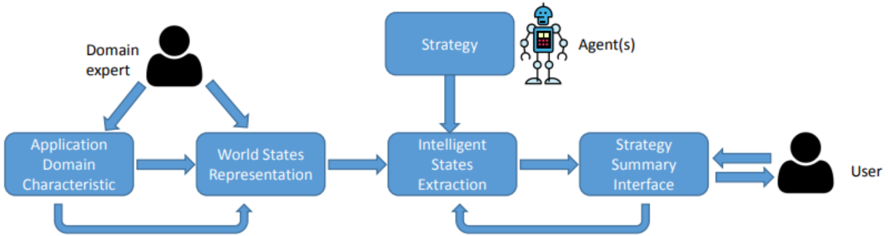
Overview of explanation in different AI fields (7)

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

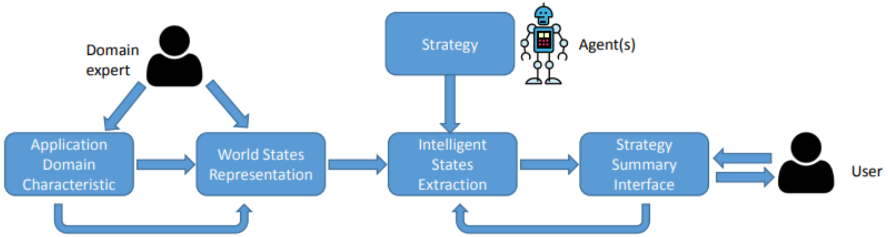
Overview of explanation in different AI fields (7)

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	

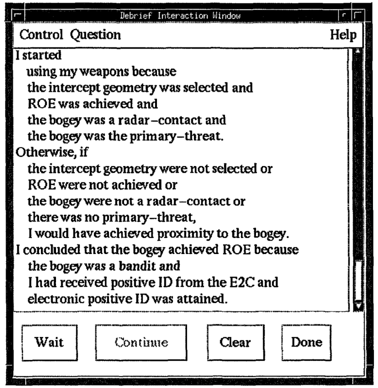
Explanation of Agent Conflicts & Harmful Interactions

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

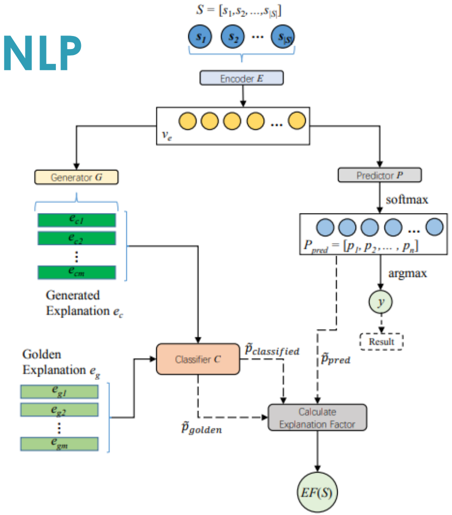
Joost Broekens, Maaike Harbers, Koen V. Hindriks, Karel van den Bosch, Catholijn M. Jonker, John-Jules Ch. Meyer: Do You Get It? User-Evaluated Explainable BDI Agents. MATES 2010: 28-39

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

THALES

Overview of explanation in different AI fields (8)

NLP



Fine-grained explanations are in the form of:

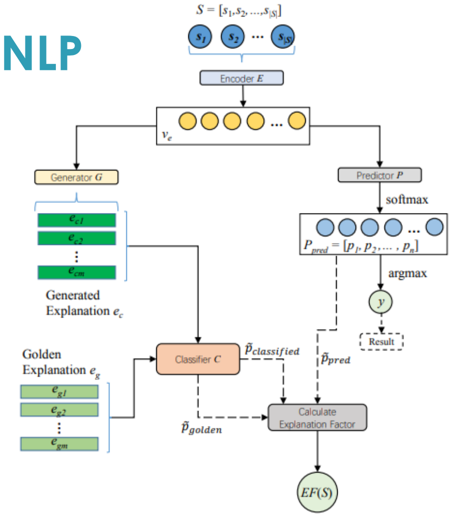
- texts in a real-world dataset;
- Numerical scores

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 (8)

NLP

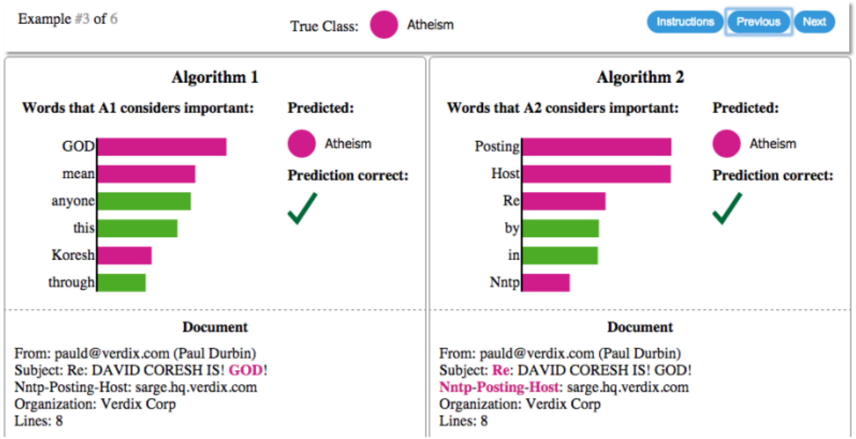


Fine-grained explanations are in the form of:

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

Explainable NLP

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

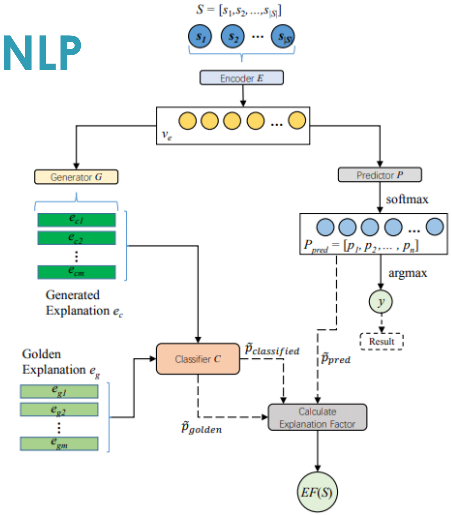


LIME for NLP

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

Overview of explanation in different AI fields (8)

NLP

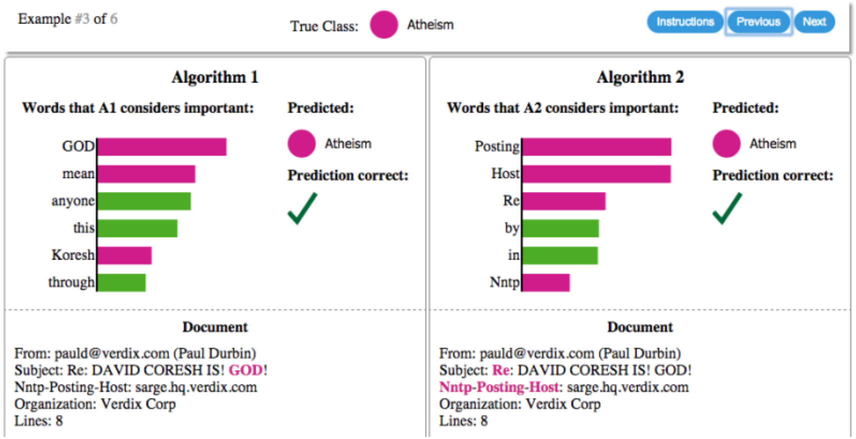


Fine-grained explanations are in the form of:

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

Explainable NLP

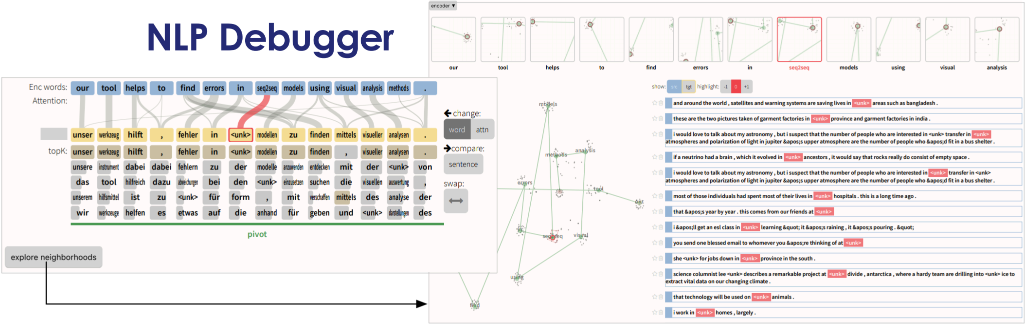
Hui Liu, Qingyu Yin, William Yang Wang: Towards Explainable NLP: A Generative Explanation Framework for Text Classification. CoRR abs/1811.00196 (2018)



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)

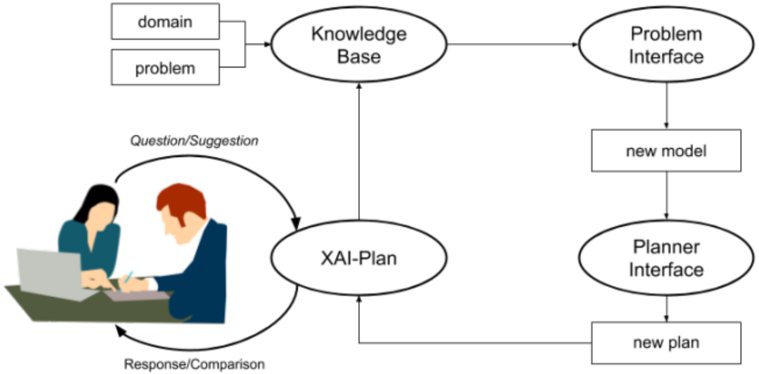
Hendrik Strobelt, Sebastian Gehrmann, Michael Behrisch, Adam Perer, Hanspeter Pfister, Alexander M. Rush: Seq2seq-Vis: A Visual Debugging Tool for Sequence-to-Sequence Models. IEEE Trans. Vis. Comput. Graph. 25(1): 353-363 (2019)

Overview of explanation in different AI fields (9)

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

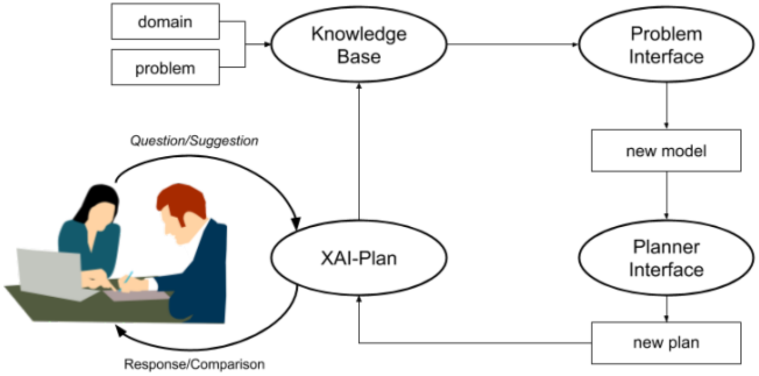
Rita Borgo, Michael Cashmore, Daniele Magazzeni: Towards Providing Explanations for AI Planner Decisions. CoRR abs/1810.06338 (2018)

Overview of explanation in different AI fields (9)

Planning and Scheduling

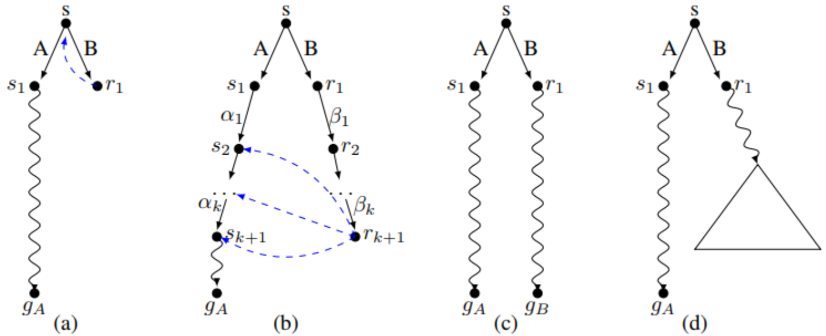
Explanation Type	R1	R2	R3	R4
Plan Patch Explanation / VAL	✗	✓	✗	✓
Model Patch Explanation	✓	✗	✓	✓
Minimally Complete Explanation	✓	✓	✗	?
Minimally Monotonic Explanation	✓	✓	✓	?
(Approximate) Minimally Complete Explanation	✗	✓	✗	✓

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)



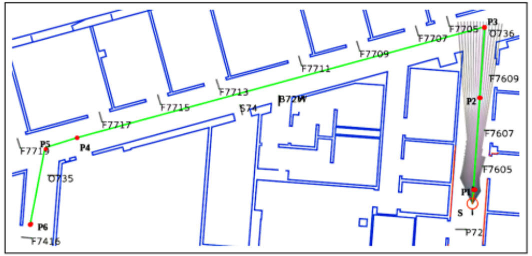
Human-in-the-loop Planning

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

(Manual) Plan Comparison

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

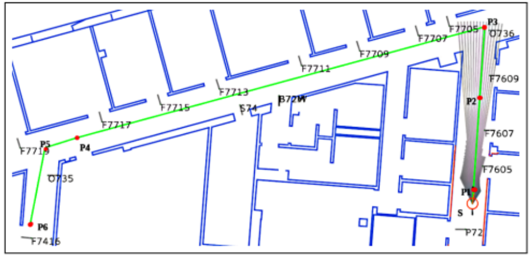
Stephanie Rosenthal, Sai P Selvaraj, and Manuela Veloso. Verbalization: Narration of autonomous robot experience. In IJCAI, pages 862–868. AAAI Press, 2016.

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

THALES

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.

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".

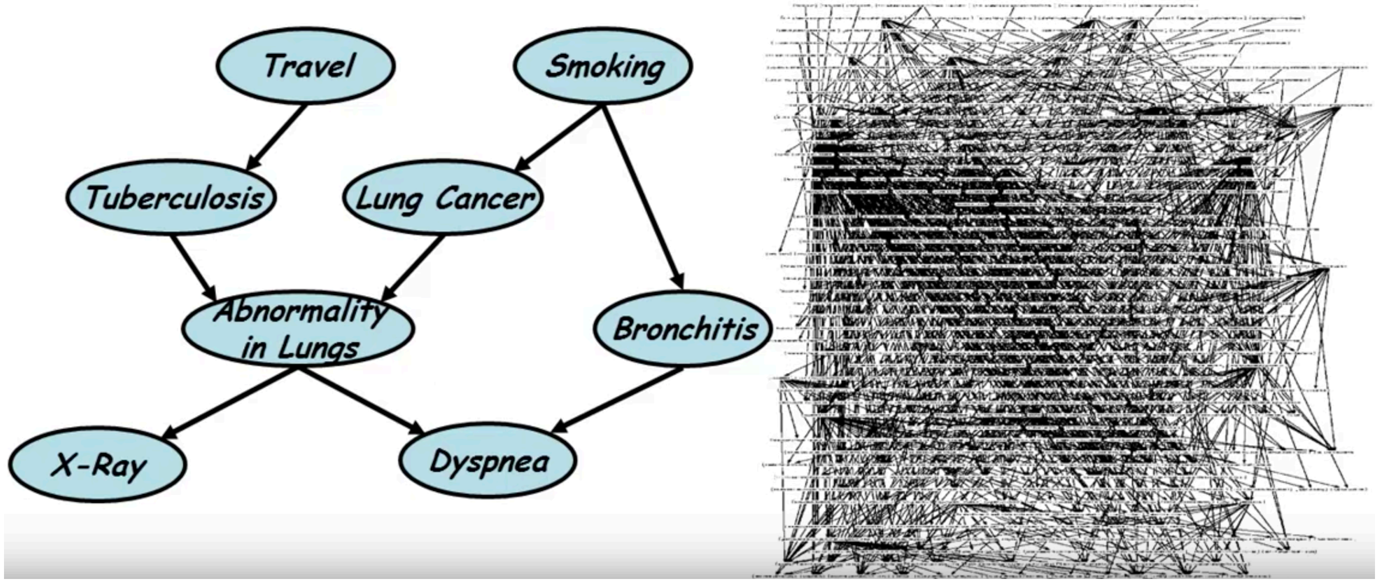
From Decision Tree to human-friendly information

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

THALES

Overview of explanation in different AI fields (11)

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

Evaluation

THALES

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

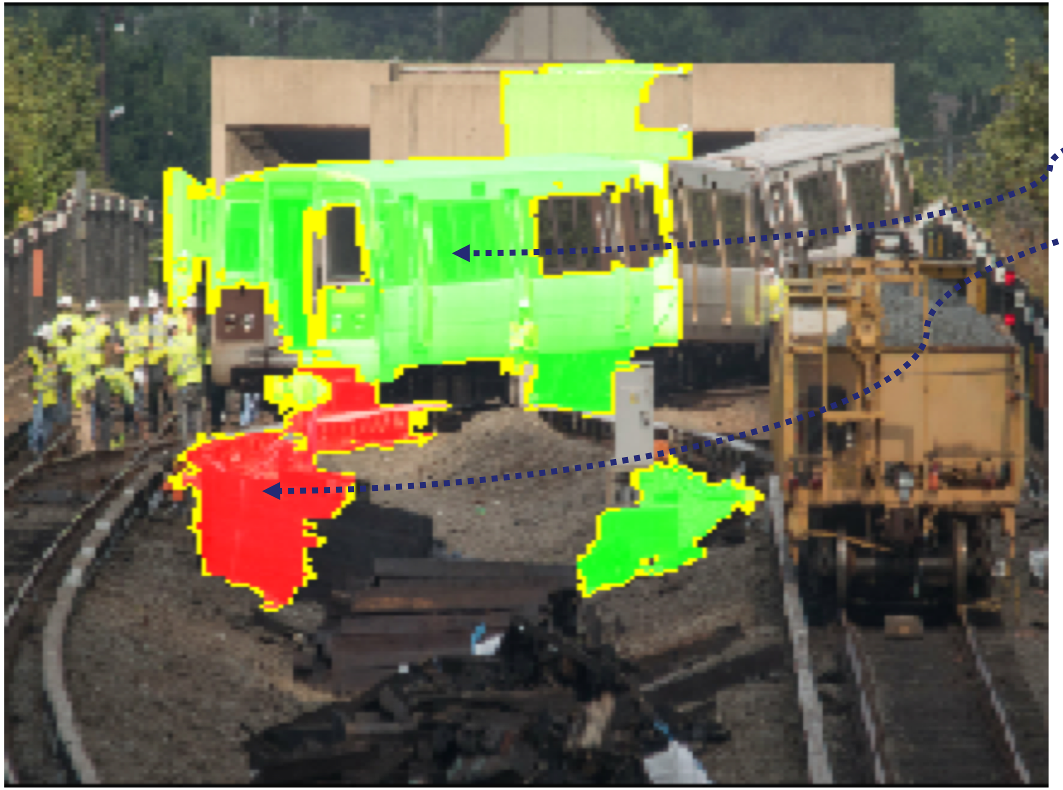
Is the explanation complete, partial, restricted?

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

On the role of Knowledge Graphs in Explainable Machine Learning

THALES

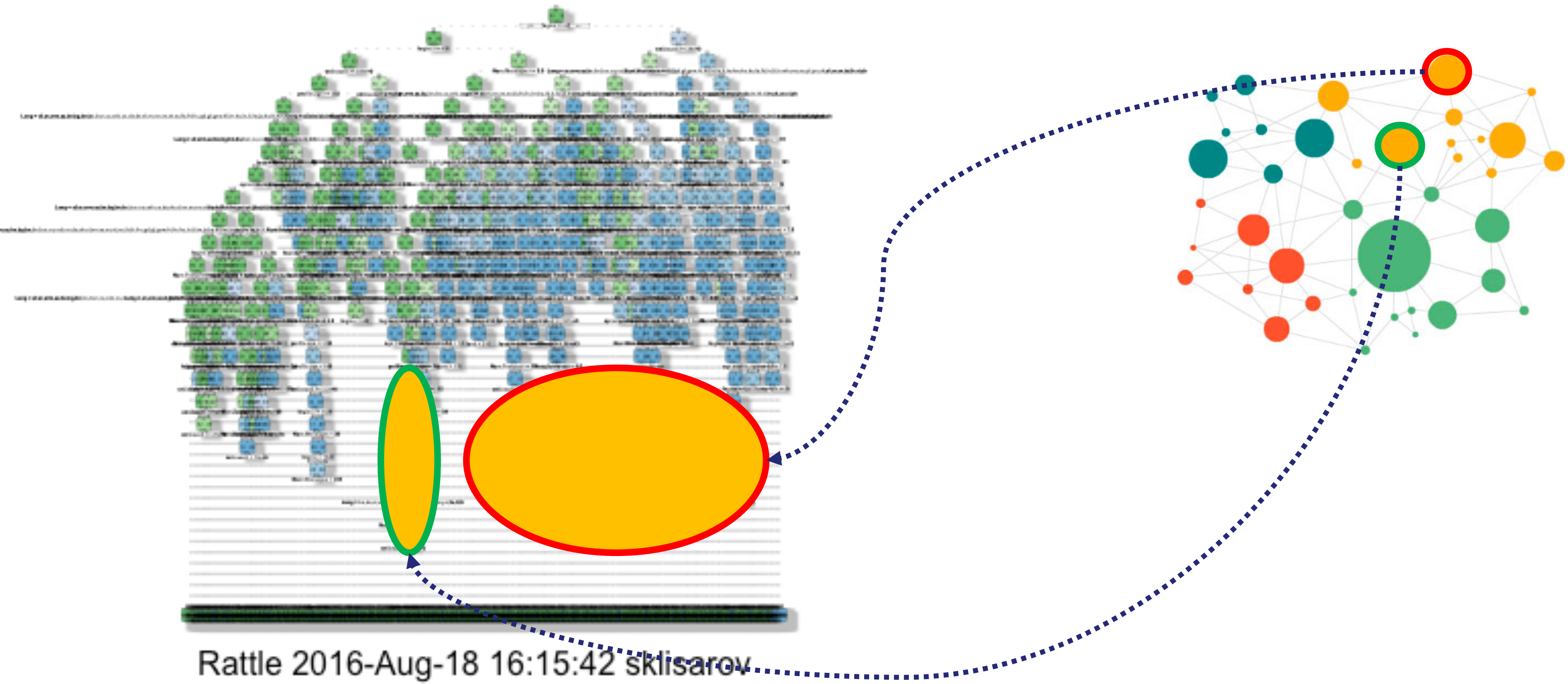
Knowledge Graph Embeddings in Machine Learning



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

THALES

Knowledge Graph for Decision Trees



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

THALES

Knowledge Graph for Deep Neural Network (1)

Input Layer

Training Data

Input (unlabeled image)

Hidden Layer

Neurons respond to simple shapes

1st Layer

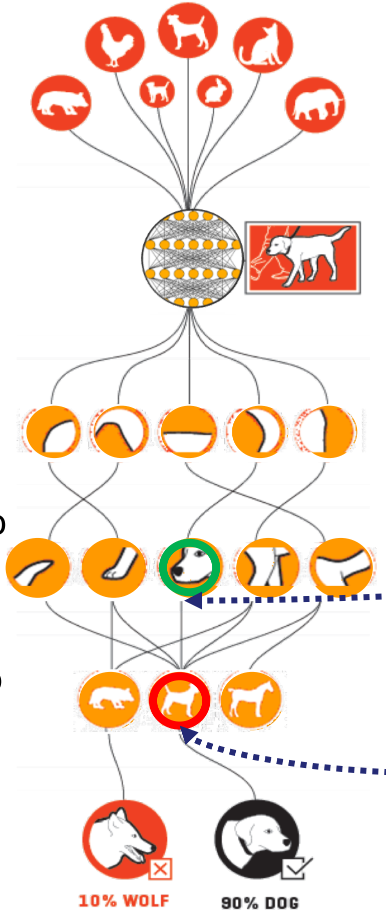
Neurons respond to more complex structures

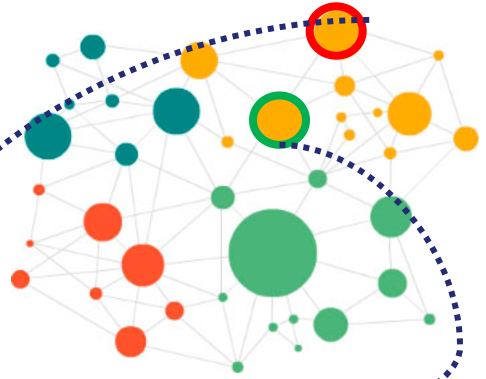
2nd Layer

Neurons respond to highly complex, abstract concepts

nth Layer

Low-level features to high-level features





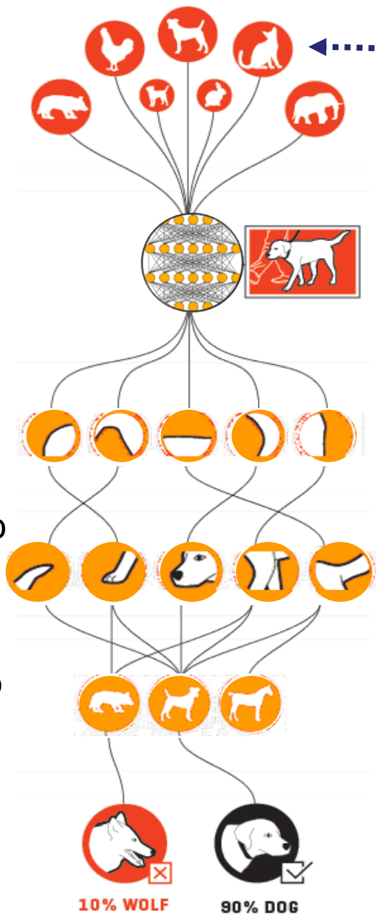
Output Layer

THALES

Knowledge Graph for Deep Neural Network (2)

Input Layer

Training Data



Input (unlabeled image)

Hidden Layer

Neurons respond to simple shapes

1st Layer

Neurons respond to more complex structures

2nd Layer

Neurons respond to highly complex, abstract concepts

nth Layer

Low-level features to high-level features

Output Layer

What is the causal relationship between the input / hidden / output layers

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

THALES

"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

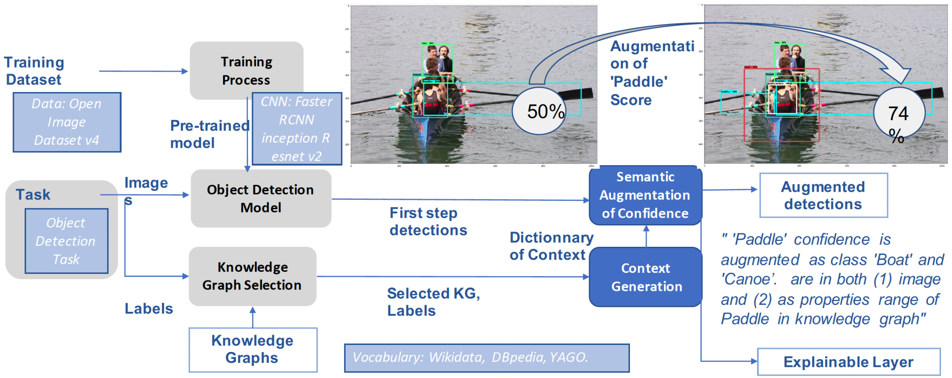
Ian Horrocks
Department of Computer Science
University of Oxford, UK

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

THALES

Applications

Explainable Boosted Object Detection – Industry Agnostic



Challenge: Object detection is usually performed from a large portfolio of Artificial Neural Networks (ANNs) architectures trained on large amount of labelled data. Explaining object detections is rather difficult due to the high complexity of the most accurate ANNs.

AI Technology: Integration of AI related technologies i.e., Machine Learning (Deep Learning / CNNs), and knowledge graphs / linked open data.

XAI Technology: Knowledge graphs and Artificial Neural Networks

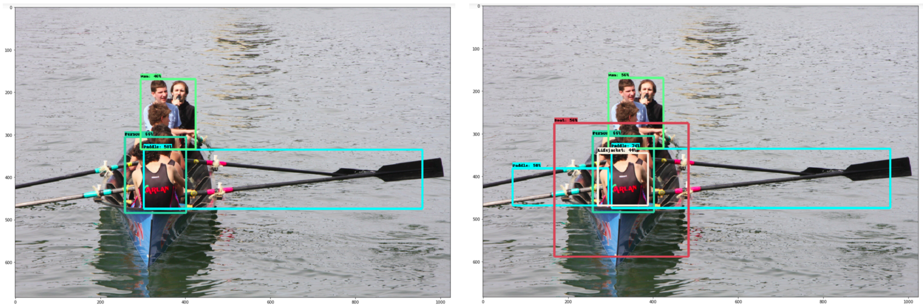
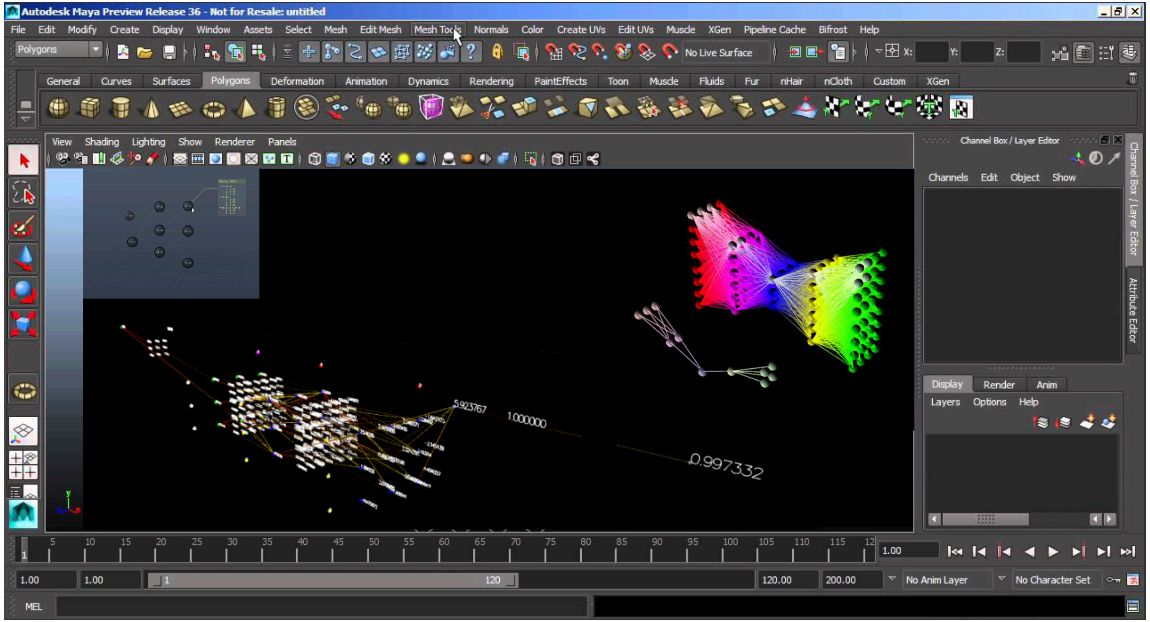


Fig. 2. Left image: results from baseline Faster RCNN: Paddle: 50% confidence, Person: 66%, Man: 46%. Right image: results from the semantic augmentation: **Paddle: 74%** confidence, Person: 66%, **Man: 56%**, **Boat: 58% with explanation: Person, Paddle, Water as part of the context in the image and knowledge graph of concept Boat**. (color print).

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

THALES

Explaining Visual Question Answering – Industry Agnostic

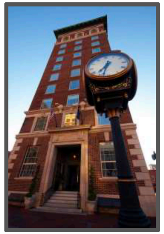
Tabular QA

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

THALES

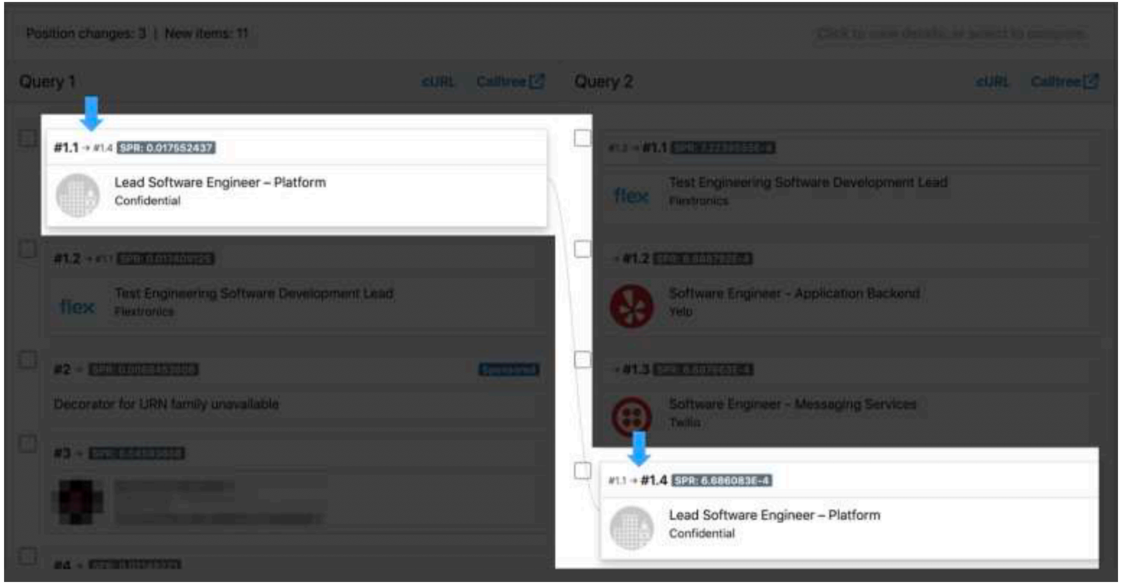
Source: Explainable AI in Industry. KDD 2019 Tutorial. Ankur Taly, Mukund Sundararajan, Kedar Dhamdhere, Pramod Mudrakarta

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?

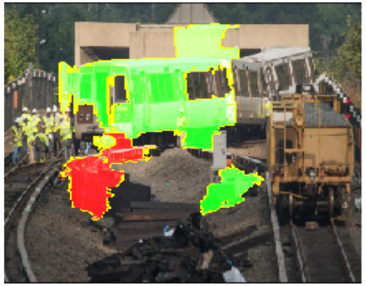
AI Technology: Artificial Neural Networks.

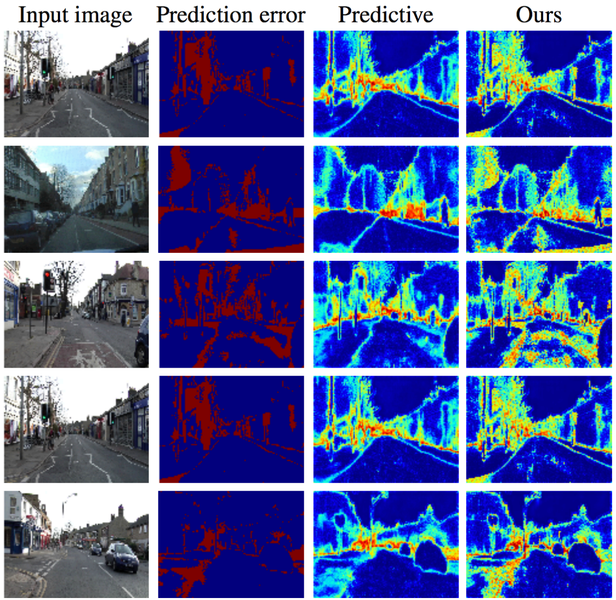
XAI Technology: Model / Prediction comparison



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

Obstacle Identification Certification (Trust) - Transportation





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

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

XAI Technology: Deep learning and Epistemic uncertainty

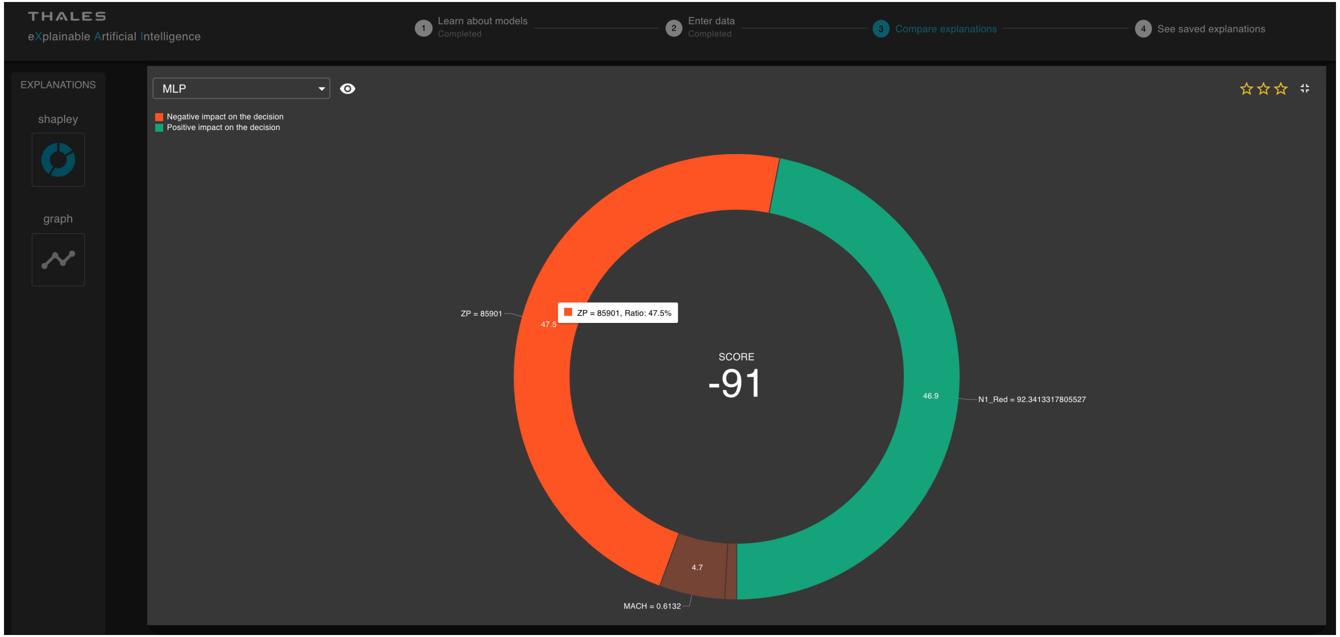
THALES

Explaining Flight Performance- Transportation

Challenge: Predicting and explaining aircraft engine performance

AI Technology: Artificial Neural Networks

XAI Technology: Shapely Values



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	-
jdsfew	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

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

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

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

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

XAI Technology: Knowledge graph embedded Sequence Learning using LSTMs

THALES

INNOVATION ARCHITECTURE:
ACCENTURE
LABS

THALES

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

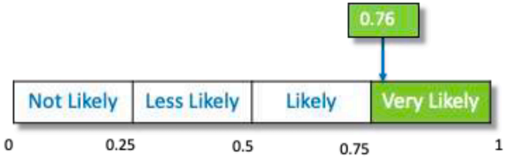
Challenge: How to predict and explain upsell / churn for a company?

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)

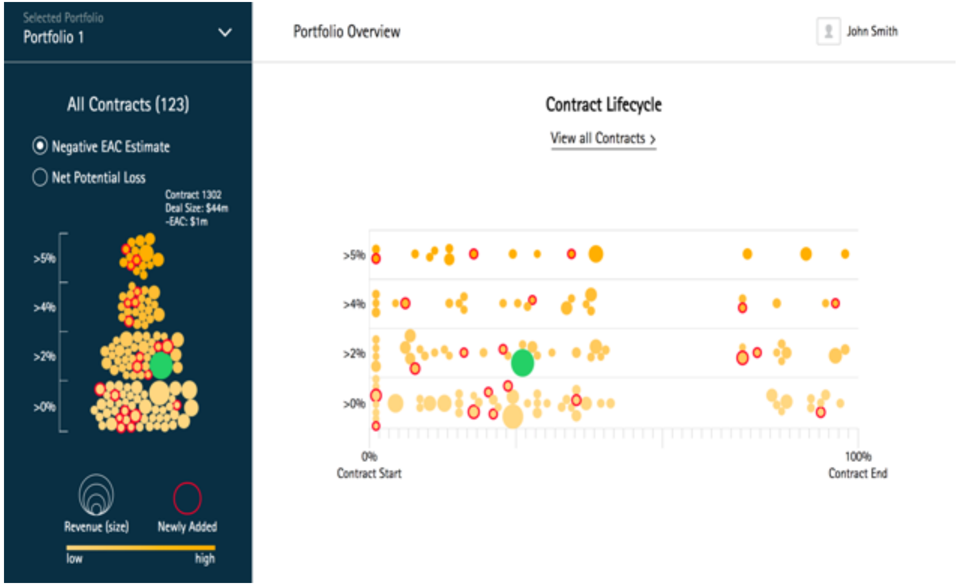
f5: 0 → 5.4, 0.03
f6: 168 → 0, 0.03
f7: 0 → 0.24, 0.02

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

INNOVATION ARCHITECTURE: ACCENTURE LABS

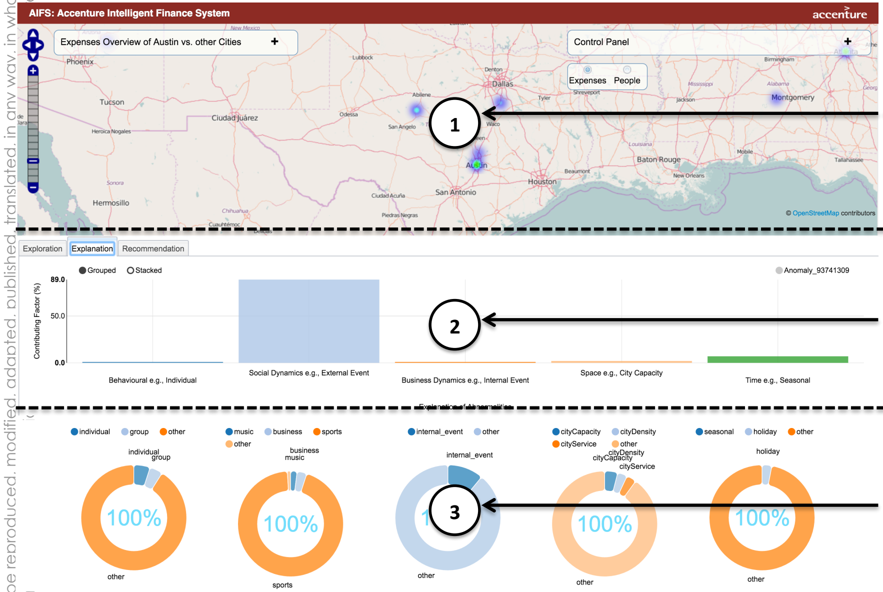
Challenge: Accenture is managing every year more than 80,000 opportunities and 35,000 contracts with an expected revenue of $34.1 billion. Revenue expectation does not meet estimation due to the complexity and risks of critical contracts. This is, in part, due to the (1) large volume of projects to assess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

THALES

Explainable Anomaly Detection – Finance (Compliance)

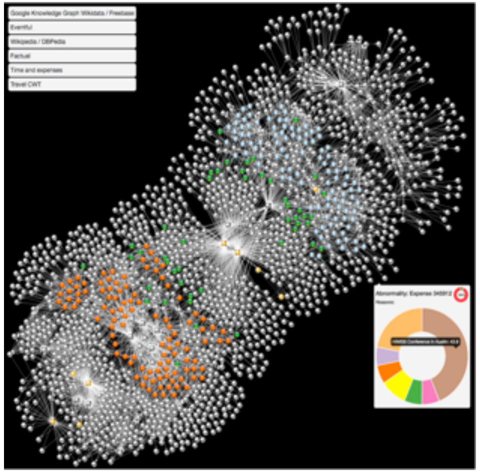


Data analysis for spatial interpretation of abnormalities: abnormal expenses

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

THALES

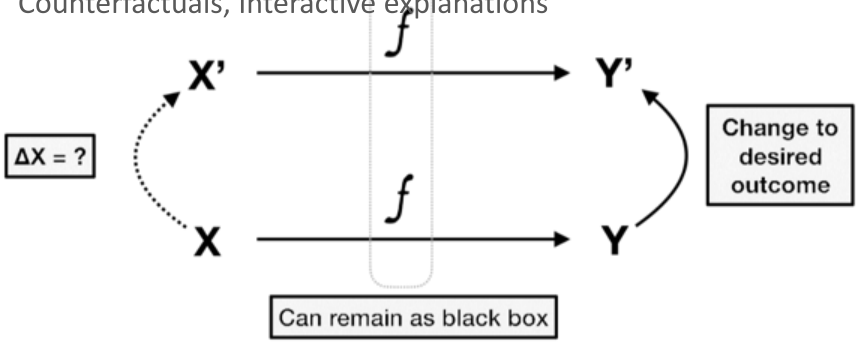
Counterfactual Explanations for Credit Decisions (1) - Finance

- **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**

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.

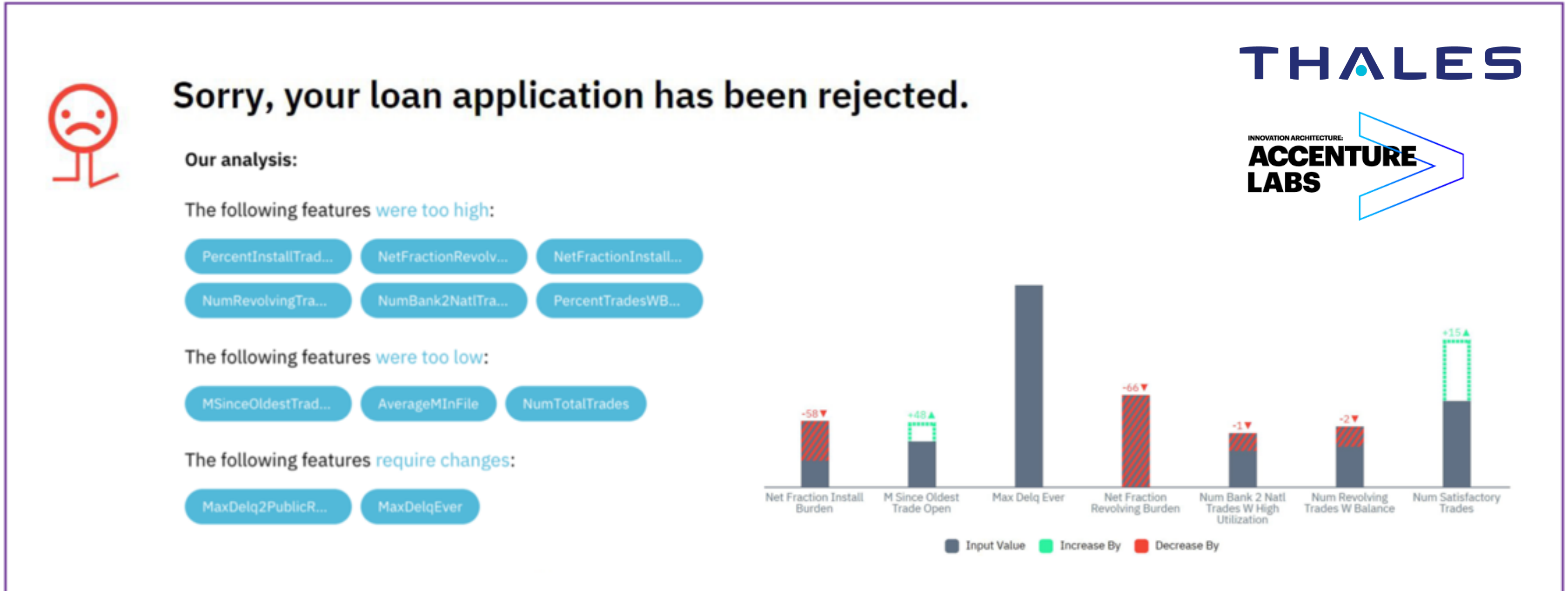
AI Technology: Supervised learning, binary classification.

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.

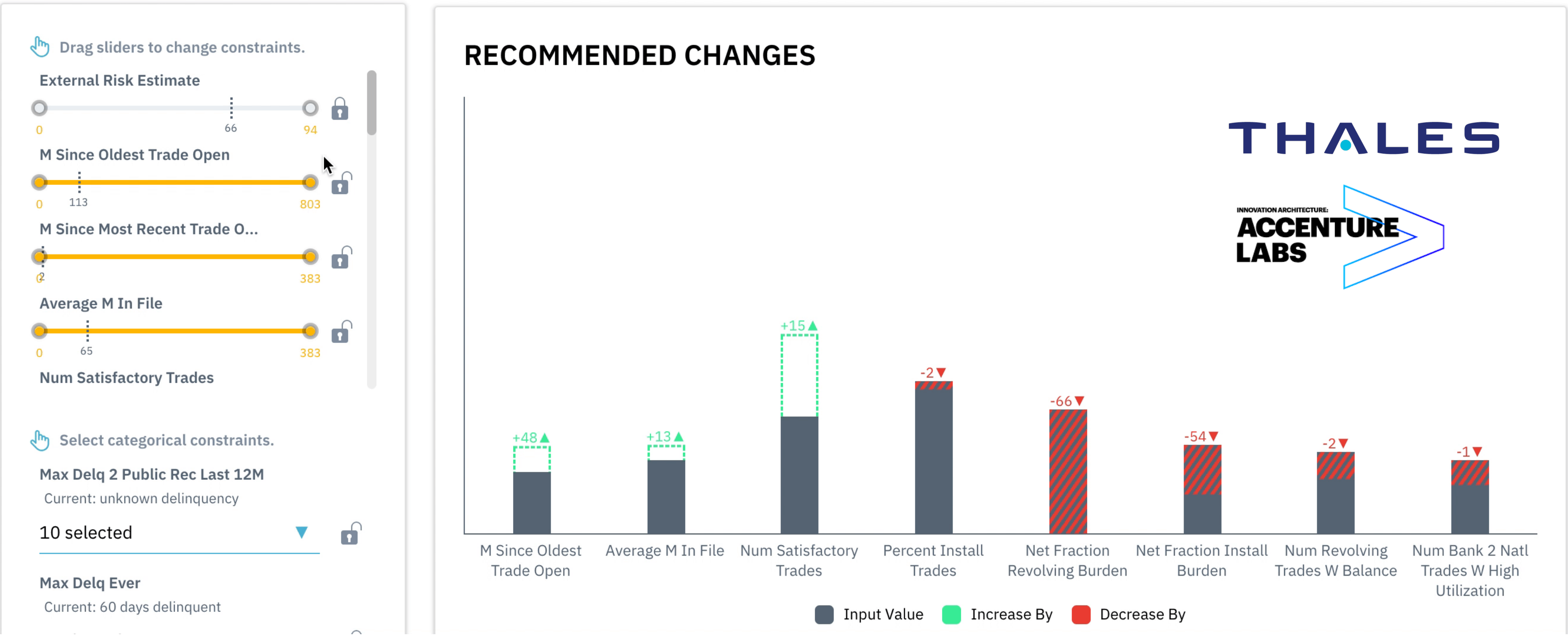
Counterfactual Explanations for Credit Decisions (2) - Finance



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

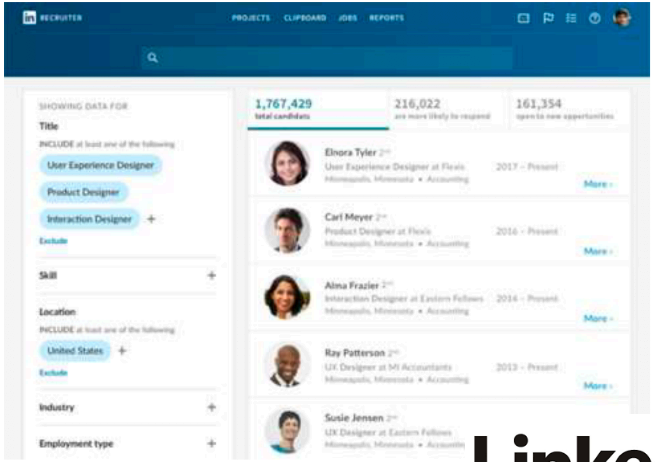
Rory Mc Grath, Luca Costabello, Chan Le Van, Paul Sweeney, Farbod Kamiab, Zhao Shen, Freddy Lécué: Interpretable Credit Application Predictions With Counterfactual Explanations. FEAP-AI4fin workshop, NeurIPS, 2018.

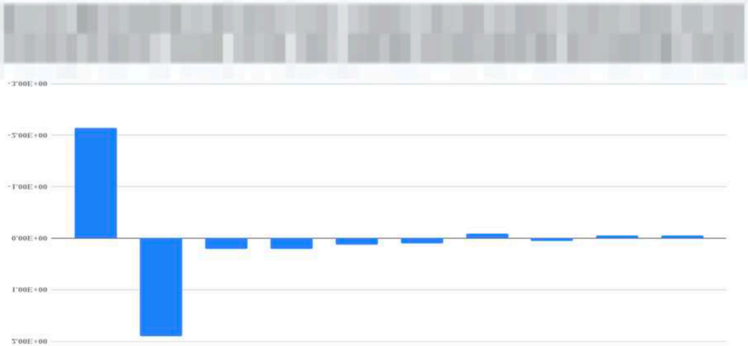
Counterfactual Explanations for Credit Decisions (3) - Finance



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





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

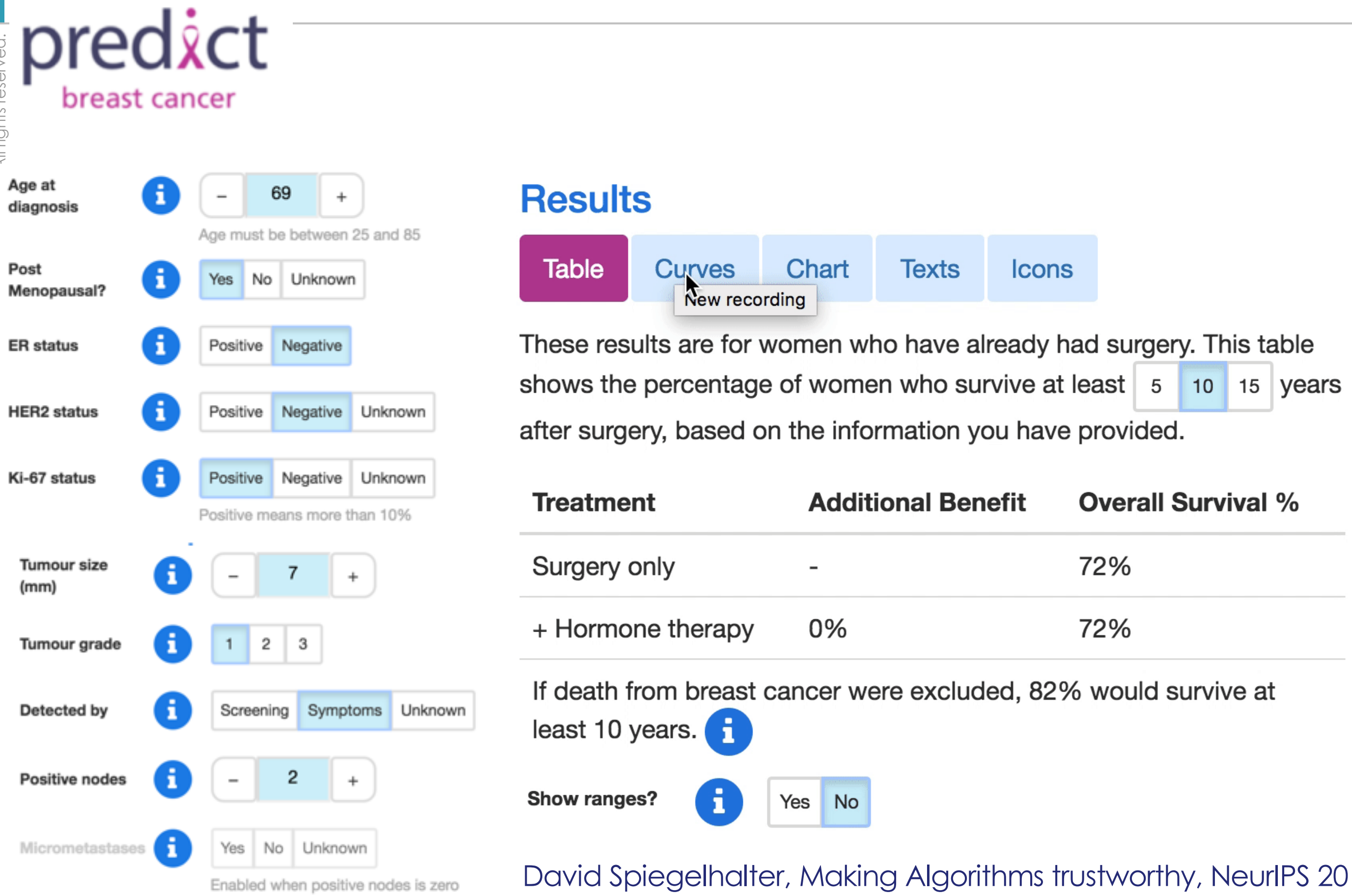
Challenge: Explaining medical condition relapse in the context of oncology.

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



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

AI Technology: competing risk analysis

XAI Technology: Interactive explanations, Multiple representations.

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

More on XAI

THALES

(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/

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

- 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/

Challenge:

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

THALES

(Some) Software Resources

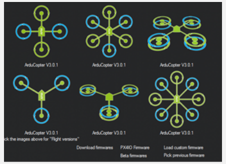
- DeepExplain: perturbation and gradient-based attribution methods for Deep Neural Networks interpretability. github.com/marcoancona/DeepExplain
- iNNvestigate: A toolbox to iNNvestigate neural networks' predictions. github.com/albermax/innvestigate
- SHAP: SHapley Additive exPlanations. github.com/slundberg/shap
- 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)*

THALES

(Some) Initiatives: XAI in USA

Challenge Problem Areas

Data Analytics
Multimedia Data

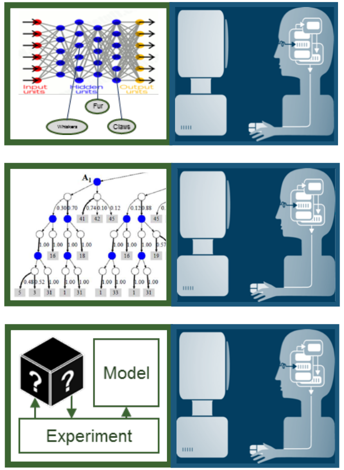


Autonomy
ArduPilot &
SITL Simulation

TA 1: Explainable Learners

Teams that provide prototype systems with both components:

- Explainable Model
- Explanation Interface



Deep Learning Teams

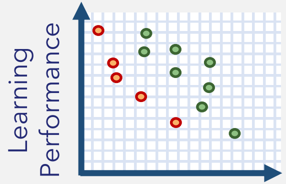
Interpretable Model Teams

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 (Dependable Explainable Learning) Project 2019-2024

- Research institutions
- Industrial partners
- Academic partners
  - Science and technology to develop new methods towards 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

- 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**

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.

hnology leader for the Defen
ogy, 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, Avio

This is a unique opportunity to play a key role on
Technology (TRT) in Canada (Quebec and Montre
applied R&T experts at five locations worldwide.
intelligence technologies. Our passion is imaginin
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 impleme
individual will also be involved in the initial projec
thinking, and team work is also critical for this rol

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

MAY 2ND, 2019

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

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