Explainable AI - XAI

From Theory to Motivation, Industrial Applications and Coding Practices

Freddy Lecue (@freddylecue)
Fosca Giannotti, Riccardo Guidotti
Pasquale Minervini

THALES UCL Inria INVENTEURS DU MONDE NUMÉRIQUE

https://github.com/flecue/xai-aaai2021

Feb. 3rd, 2021

https://xaitutorial2021.github.io

Outline

Agenda

- Part I: Introduction, Motivation & Evaluation *– 20 minutes*
    - Motivation, Definitions & Properties
    - Evaluation Protocols & Metrics
- Part II: Explanation in AI (not only Machine Learning!) *– 40 minutes*
    - From Machine Learning to Knowledge Representation and Reasoning and Beyond
- Part III: On The Role of Knowledge Graphs in Explainable Machine Learning *– 40 minutes*
- Part IV: XAI Tools and Coding Practices – 40 minutes
- Part V: Applications, Lessons Learnt and Research Challenges – 40 minutes
    - Explaining (1) object detection, (2) obstacle detection for autonomous trains, (3) flight performance, (4) flight delay prediction, (5) risk management, (6) abnormal expenses, (7) credit decisions, (8) medical conditions + 8 more use cases in industry

Scope

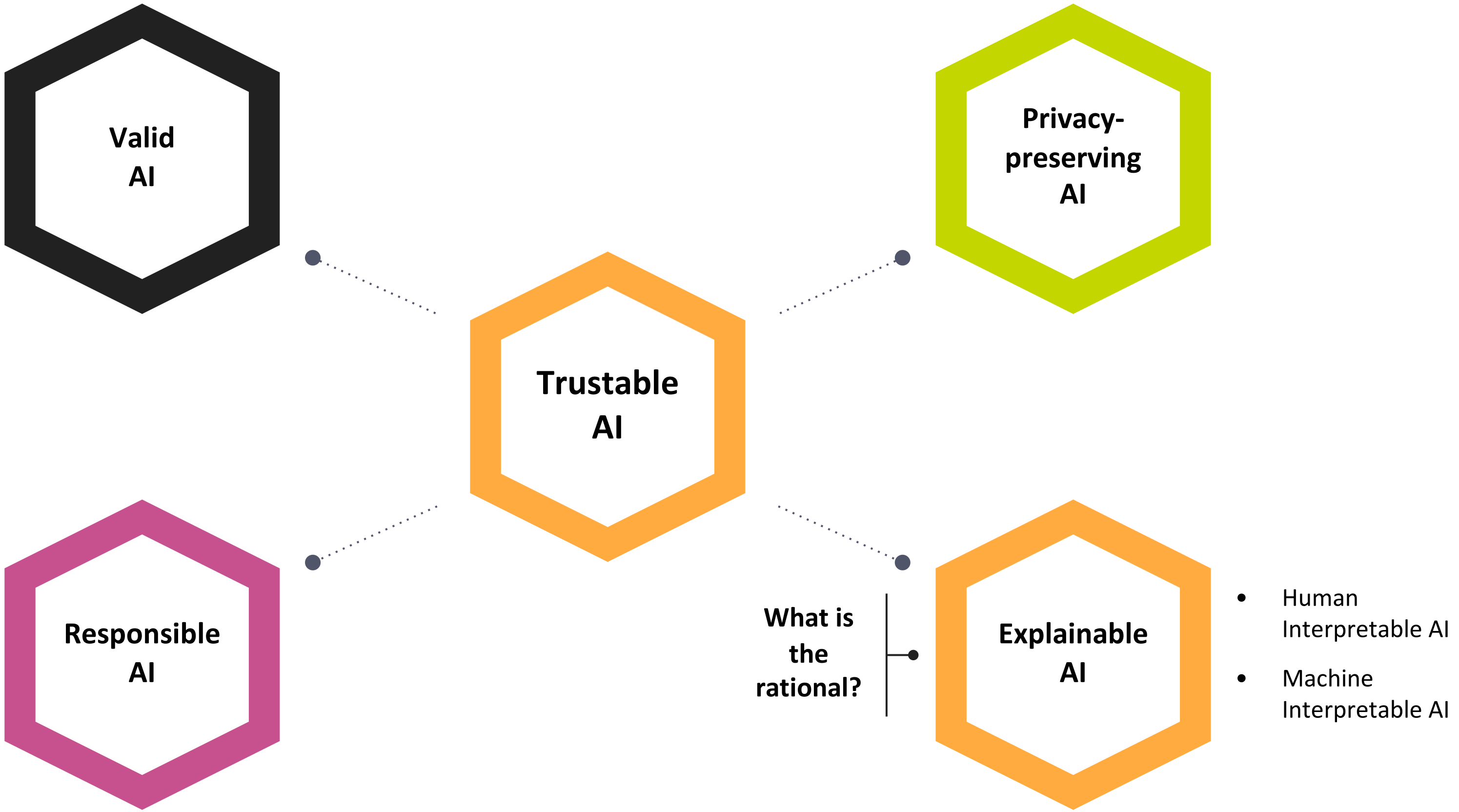
Disclaimer

- **As MANY interpretations as research areas**

(check out work in Machine Learning vs Reasoning community)

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

AI Adoption: Requirements



Explainability **Fairness** **Privacy** **Transparency**

SR 11-7: Guidance on Model Risk Management

BOARD OF GOVERNORS
OF THE FEDERAL RESERVE SYSTEM
WASHINGTON, D.C. 20551

What's driving Stress Testing and Model Risk Management efforts?

Regulatory efforts

SR 11-7 says "Banks benefit from **conducting model stress testing** to check performance over a wide range of inputs and parameter values, including extreme values, **to verify that the model is robust**"

In fact, **SR14-03** explicitly calls for **all models used for Dodd-Frank Act Company-Run Stress Tests must fall under the purview of Model Risk Management**.

In addition **SR12-07** calls for **incorporating validation or other type of independent review of the stress testing framework to ensure the integrity of stress testing processes and results.**

CALIFORNIA CONSUMER PRIVACY ACT OF 2018

GDPR

Article 22. Automated individual decision making, including profiling

1. The data subject shall have the right not to be subject to a decision based solely on automated processing, including profiling, which produces legal effects concerning him or her or similarly significantly affects him or her.
2. Paragraph 1 shall not apply if the decision:
   (a) is necessary for entering into, or performance of, a contract between the data subject and a data controller;
   (b) is authorised by Union or Member State law to which the controller is subject and which also lays down suitable measures to safeguard the data subject's rights and freedoms and legitimate interests; or
   (c) is based on the data subject's explicit consent.
3. In the cases referred to in points (a) and (c) of paragraph 2, the data controller shall implement suitable measures to safeguard the data subject's rights and freedoms and legitimate interests, at least the right to obtain human intervention on the part of the controller, to express his or her point of view and to contest the decision.
4. Decisions referred to in paragraph 2 shall not be based on special categories of personal data referred to in Article 9(1), unless point (a) or (g) of Article 9(2) apply and suitable measures to safeguard the data subject's rights and freedoms and legitimate interests are in place.

Credit: Lecue et al., Tutorial on XAI. AAAI 2020. https://xaitutorial2020.github.io/

Growing Global AI Regulation

- GDPR: Article 22 empowers individuals with the right to demand an explanation of how an automated system made a decision that affects them.
- Algorithmic Accountability Act 2019: Requires companies to **provide an assessment of the risks** posed by the automated decision system to the **privacy** or **security** and the risks that contribute to **inaccurate, unfair, biased, or discriminatory decisions** impacting consumers
- California Consumer Privacy Act: Requires companies to rethink their approach to capturing, storing, and sharing personal data to align with the new requirements by January 1, 2020.
- Washington Bill 1655: Establishes guidelines for the use of automated decision systems to protect consumers, improve transparency, and create more market predictability.
- Massachusetts Bill H.2701: Establishes a commission on automated decision-making, transparency, fairness, and individual rights.
- Illinois House Bill 3415: States predictive data analytics determining creditworthiness or hiring decisions may not include information that correlates with the applicant race or zip code.

Credit: Lecue et al., Tutorial on XAI. AAAI 2020. https://xaitutorial2020.github.io/

Part I

Introduction and Motivation

Explanation - From a Business Perspective

Business to Customer AI

Critical Systems (1)

Critical Systems (2)

… but not only Critical Systems (1)

COMPAS recidivism black bias

… but not only Critical Systems (2)

Finance:

- Credit scoring, loan approval
- Insurance quotes

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

The Big Read **Artificial intelligence** + Add to myFT

Insurance: Robots learn the business of covering risk

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

Save

Oliver Ralph MAY 16, 2017 24

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

… but not only Critical Systems (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

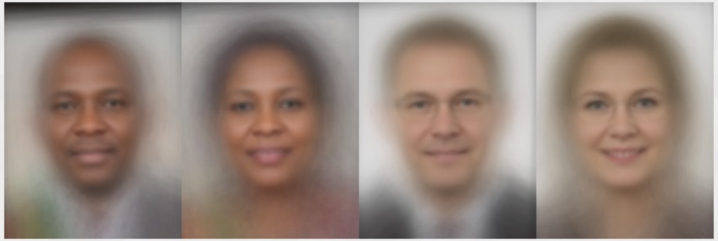
Marc Sturm
NewYork-Presbyterian Hospital
mas9161@nyp.org

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

… and even More

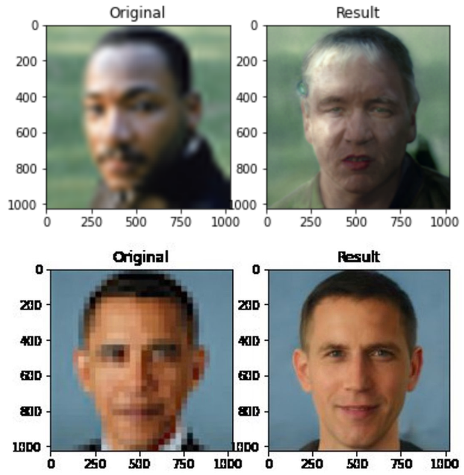
Gender Classifier	Darker Male	Darker Female	Lighter Male	Lighter Female	Largest Gap
Microsoft	94.0%	79.2%	100%	98.3%	20.8%
FACE++	99.3%	65.5%	99.2%	94.0%	33.8%
IBM	88.0%	65.3%	99.7%	92.9%	34.4%



Joy Buolamwini, Timnit Gebru: Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. FAT 2018: 77-91

https://techcrunch.com/2020/10/02/twitter-may-let-users-choose-how-to-crop-image-previews-after-bias-scrutiny/

https://www.cbsnews.com/news/apple-credit-card-goldman-sachs-disputes-claims-that-apple-card-is-sexist/



https://www.theverge.com/21298762/face-depixelizer-ai-machine-learning-tool-pulse-stylegan-obama-bias

Explanation - In a Nutshell

XAI Definitions - Explanation vs. Interpretation

Oxford Dictionary of English

explanation | ɛkspləˈneɪʃ(ə)n |

noun

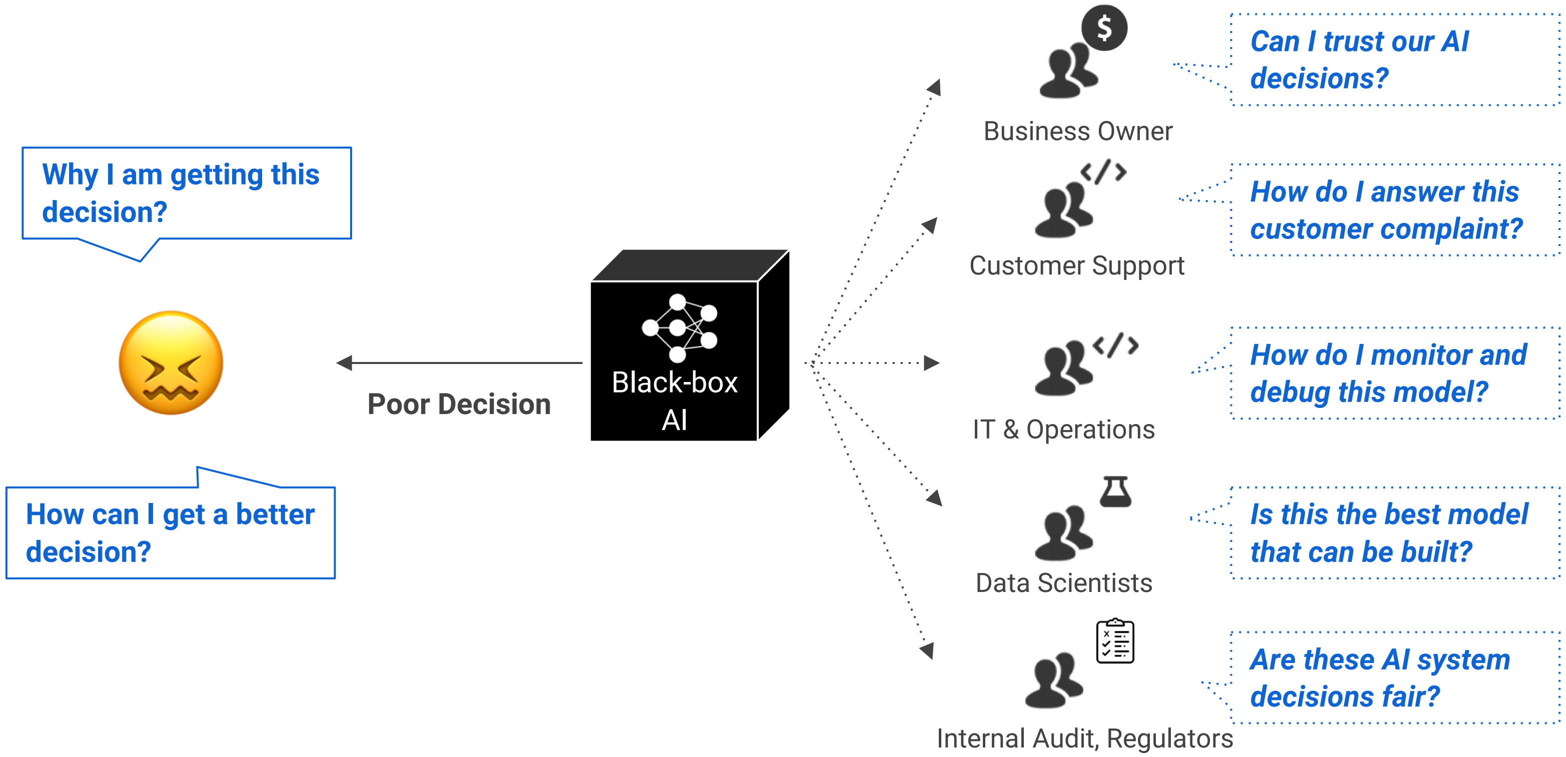
a statement or account that makes something clear: *the birth rate is central to any explanation of population trends.*

interpret | ɪnˈtəːprɪt |

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

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

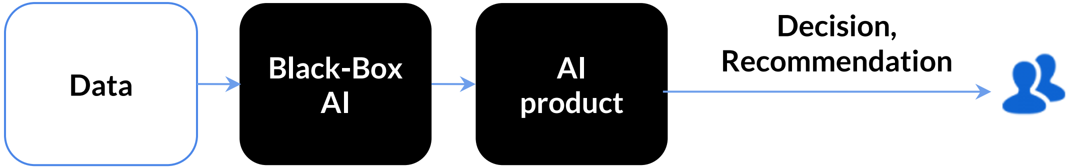
AI as a Black-box: Source of Confusion and Doubt



Credit: Lecue et al., Tutorial on XAI. AAAI 2020. https://xaitutorial2020.github.io/

XAI

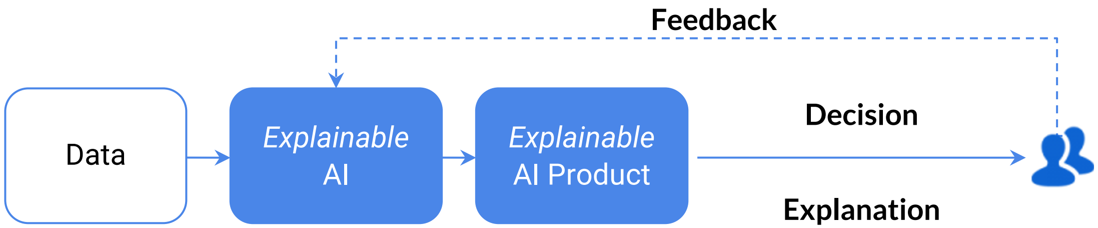
Black Box AI



Confusion with Today's AI Black Box

- Why did you do that?
- Why did you not do that?
- When do you succeed or fail?
- How do I correct an error?

Explainable AI

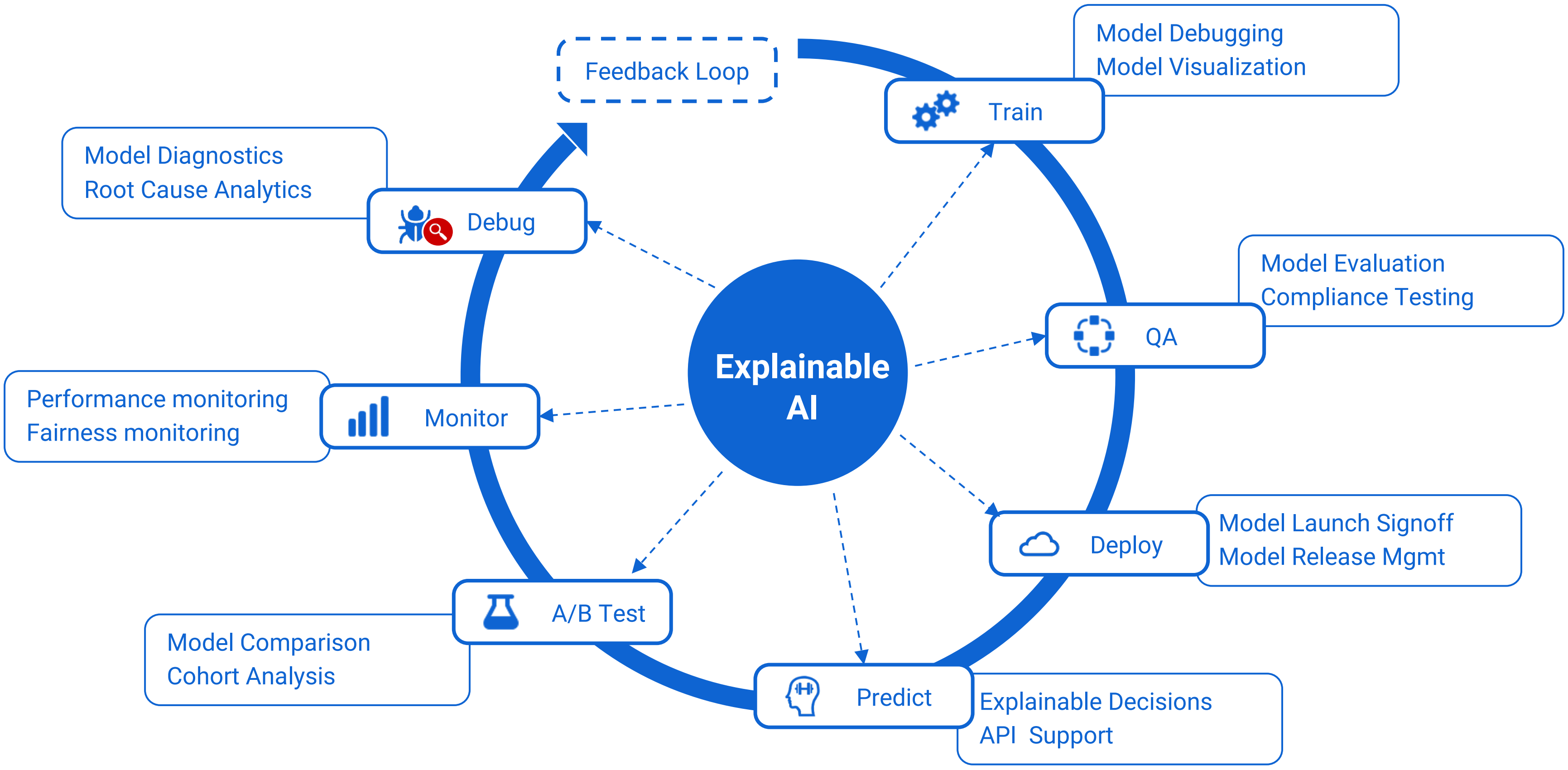


Clear & Transparent Predictions

- I understand why
- I understand why not
- I know why you succeed or fail
- I understand, so I trust you

Credit: Lecue et al., Tutorial on XAI. AAAI 2020. https://xaitutorial2020.github.io/

Explainability by Design for AI products

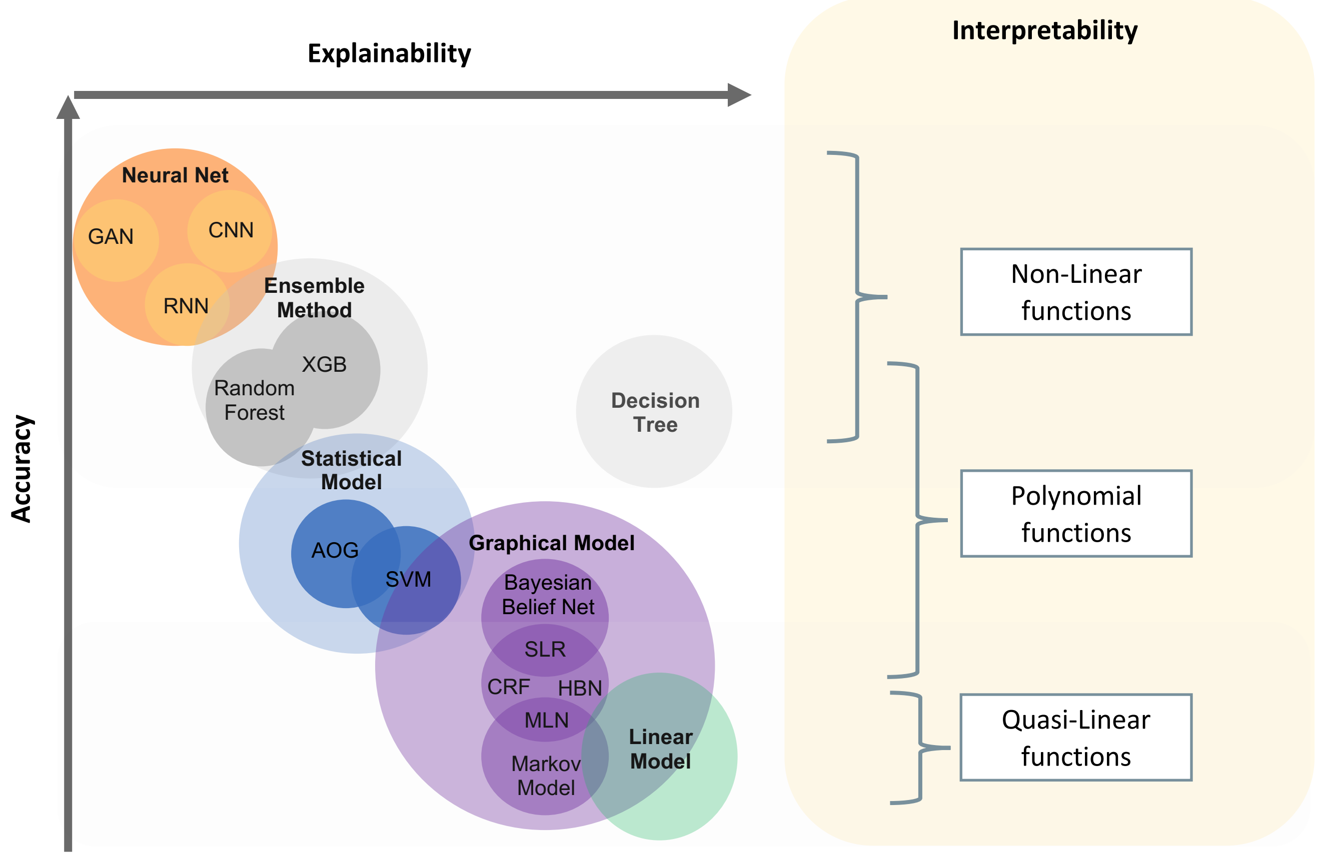


Credit: Lecue et al., Tutorial on XAI. AAAI 2020. https://xaitutorial2020.github.io/

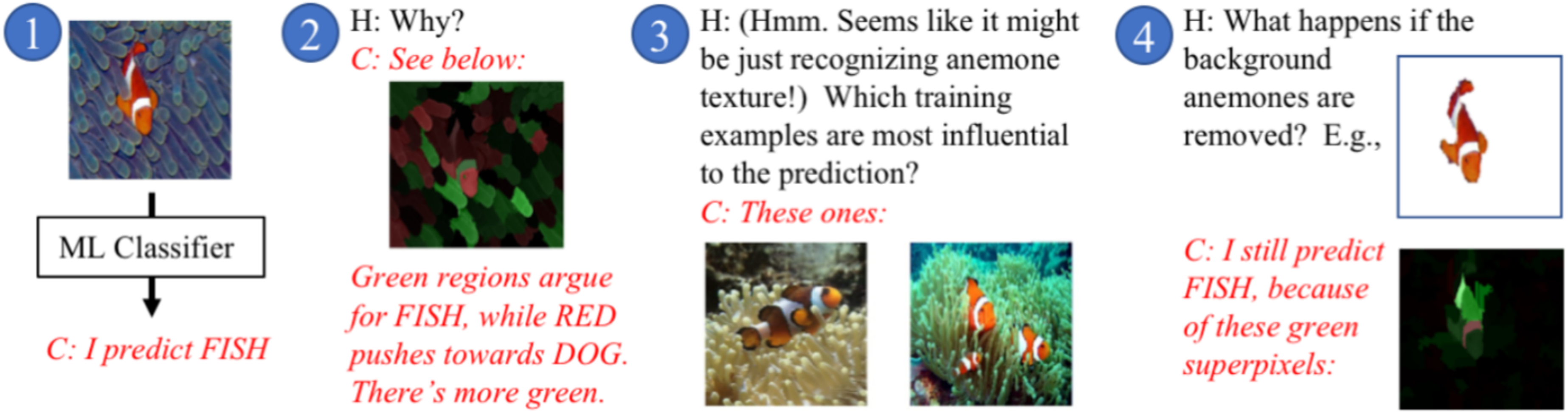
How to Explain? Accuracy vs. Explainability

Learning

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



Example of an End-to-End XAI System



- Humans may have follow-up questions
- Human – Machine interactions are required
- Explanations cannot answer all users' concerns in one shot
  - Many different stakeholders
  - Many different objectives
  - Many different expertise

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

On the Role of Data in XAI

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

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

Field names

One row (4 fields)

2000 rows all told

Tabular

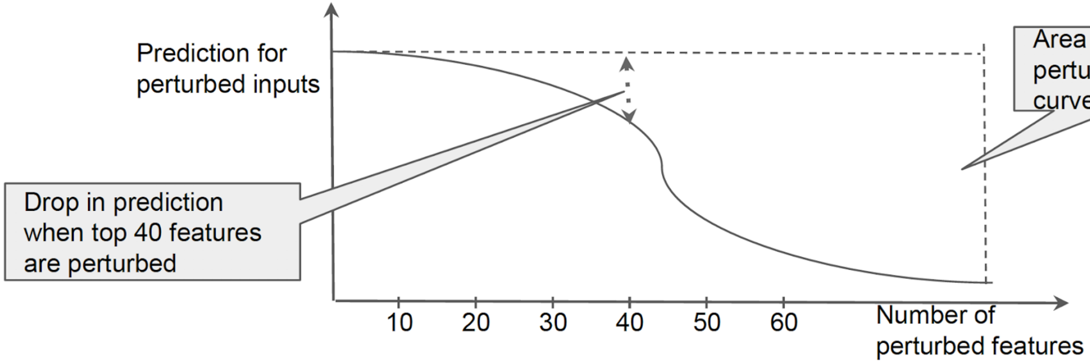
Images

Text

Evaluation (1) - Perturbation-based Approaches

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

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



Evaluation (2) – From size-based to **Human (Role)-based Evaluation**

Evaluation criteria for Explanations [Miller, 2017]
- Truth & probability
- Usefulness, relevance
- Coherence with prior belief
- Generalization

Finale Doshi-Velez, Been Kim: A Roadmap for a Rigorous Science of Interpretability. CoRR abs/1702.08608 (2017)

Forough Poursabzi-Sangdeh, Daniel G. Goldstein, Jake M. Hofman, Jennifer Wortman Vaughan, Hanna M. Wallach: Manipulating and Measuring Model Interpretability. CoRR abs/1802.07810 (2018) 18]

Cognitive chunks = basic explanation units (for different explanation needs)
- Which basic units for explanations?
- How many?
- How to compose them?
- Uncertainty & end users?

Human Factors in Explanation
- Humans prefer explanations that are both simple and highly probable
- Humans appeal to causal structure and counterfactual
- Larger explanations might push humans into a more careful, rational thinking mode.

Frank Keil. Explanation and understanding.Annu. Rev. Psychol., 2006.

Tania Lombrozo. The structure and function of explanations.Trends in cognitive sciences, 10(10):464–470, 2006.

Isaac Lage, Emily Chen, Jeffrey He, Menaka Narayanan, Been Kim, Sam Gershman, Finale Doshi-Velez: An Evaluation of the Human-Interpretability of Explanation. CoRR abs/1902.00006 (2019)

Daniel Kahneman.Thinking, fast and slow. Macmillan, 2011.

A/B Testing for Interpretable ML
- Performance on a classification task was better when using examples as representation than when using non-example-based representation
- Subjects are faster and more accurate at describing local decision boundaries based on decision sets rather than rule lists

B. Kim, C. Rudin, and J.A. Shah. The Bayesian Case Model: A generative approach for case-based reasoning and prototype classification. In NIPS, 2014.

Himabindu Lakkaraju, Stephen H Bach, and Jure Leskovec. Interpretable decision sets: A jointframework for description and prediction. InProceedings of the 22nd ACM SIGKDD, 2016.

Evaluation (3) – Example-based Explanation is Better Designed for Humans

Task	Image Recognition	Sentiment Analysis	Key Word Detection	Heartbeat Classification
Domain	Image	Text	Audio	Sensory data (ECG)
Dataset	Cifar-10	Sentiment140	Speech Commands	MIT-BIH Arrhythmia
Classes	10	2	10	5

Table 2: An overview of the application tasks and datasets used in our study

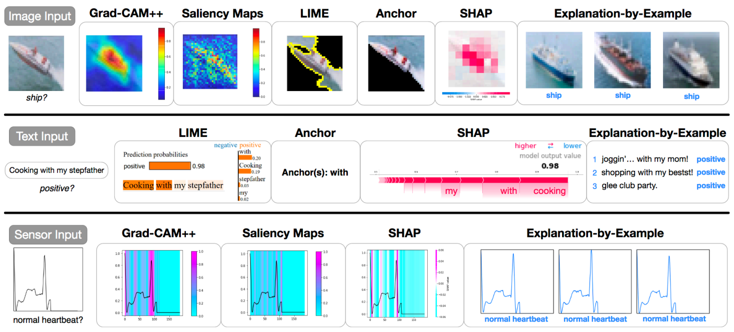
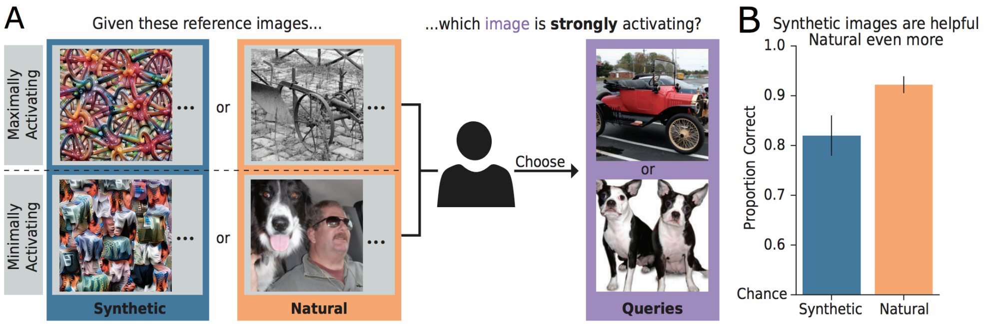


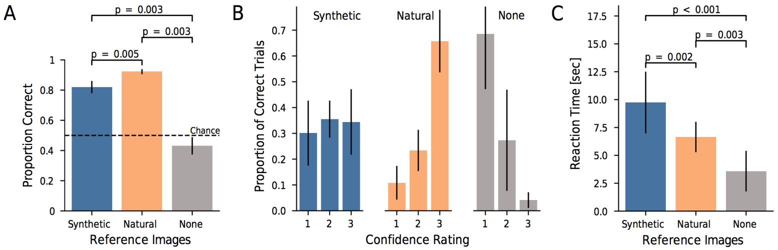
Figure 2: Depiction of surveyed explanation methods for image, text, and ECG input.

Explanation Method	Image Study	Text Study	Audio Study	Sensor Study
LIME	47.7 ± 4.5%	**70.4 ± 3.6%**	-	-
Anchor	38.9 ± 4.3%	25.8 ± 3.5%	-	-
SHAP	33.7 ± 4.3%	59.9 ± 3.8%	34.7 ± 4.8%	32.8 ± 3.3%
Saliency Maps	39.4 ± 4.3%	-	46.1 ± 5.1%	40.4 ± 3.5%
Grad-CAM++	50.8 ± 4.5%	-	48.1 ± 5.3%	42.0 ± 3.5%
ExMatchina	**89.6 ± 2.6%**	43.7 ± 3.9%	**70.9 ± 4.7%**	**84.8 ± 2.5%**

Table 3: Results of the Mechanical Turk study evaluating user preference for DNN explanation methods across image, text, audio, and sensory input domains. Survey questions individually compare two methods at a time, with each explanation compared to all other available methods equally. Results indicate the rate by which users selected a particular method when it is an available explanation, with 95% bootstrap confidence intervals.

Jeya Vikranth Jeyakumar, Joseph Noor, Yu-Hsi Cheng, Luis Garcia, Mani B. Srivastava:How Can I Explain This to You? An Empirical Study of Deep Neural Network Explanation Methods. NeurIPS 2020





Examples are (most of the time) better

Judy Borowski, Roland S. Zimmermann, Judith Schepers, Robert Geirhos, Thomas S. A. Wallis, Matthias Bethge, Wieland Brendel: Exemplary Natural Images Explain CNN Activations Better than Feature Visualizations. ICLR 2021.

Evaluation (4) – Humans Have Preferred Explanation Depending on Data

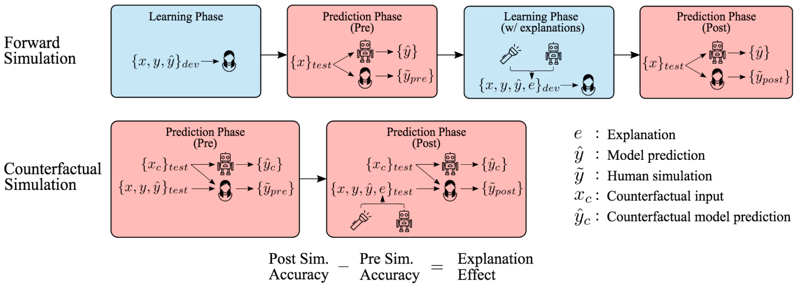


Figure 1: Forward and counterfactual simulation test procedures. We measure human users' ability to predict model behavior. We isolate the effect of explanations by first measuring baseline accuracy, then measuring accuracy after users are given access to explanations of model behavior. In the forward test, the explained examples are distinct from the test instances. In the counterfactual test, each test instance is a counterfactual version of a model input, and the explanations pertain to the original inputs.

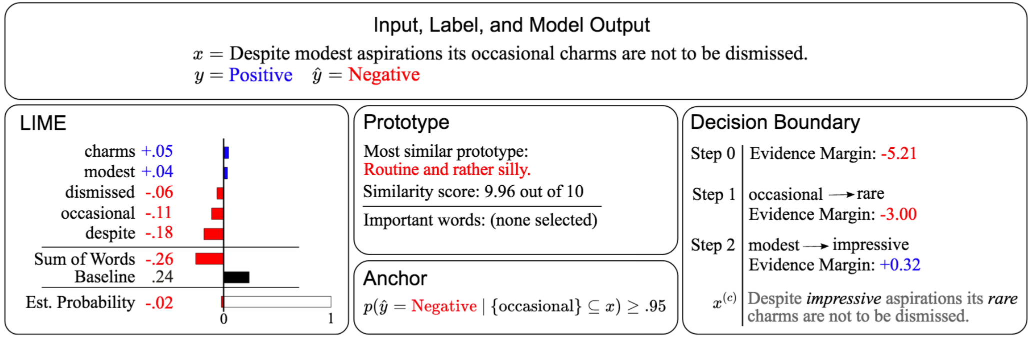


Figure 2: Explanation methods applied to an input from the test set of movie reviews.

	Text					Tabular				
Method	n	Pre	Change	CI	p	n	Pre	Change	CI	p
User Avg.	1144	62.67	-	7.07	-	1022	70.74	-	6.96	-
LIME	190	-	0.99	9.58	.834	179	-	**11.25**	8.83	.014
Anchor	181	-	1.71	9.43	.704	215	-	5.01	8.58	.234
Prototype	223	-	3.68	9.67	.421	192	-	1.68	10.07	.711
DB	230	-	−1.93	13.25	.756	182	-	5.27	10.08	.271
Composite	320	-	3.80	11.09	.486	254	-	0.33	10.30	.952

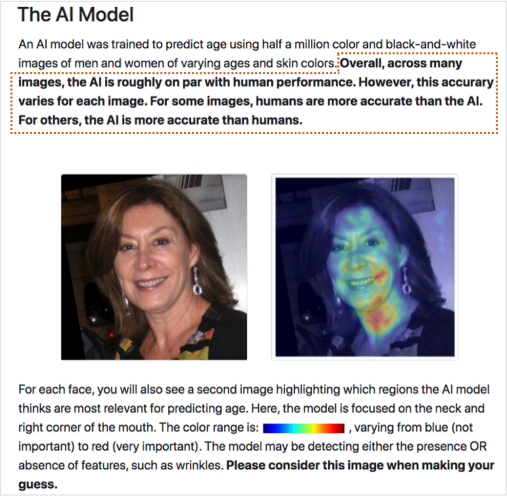
Table 1: Change in user accuracies after being given explanations of model behavior, relative to the baseline performance (Pre). Data is grouped by domain. CI gives the 95% confidence interval, calculated by bootstrap using n user responses, and we bold results that are significant at a level of $p < .05$. LIME improves simulatability with tabular data. Other methods do not definitively improve simulatability in either domain.

	Forward Simulation					Counterfactual Simulation				
Method	n	Pre	Change	CI	p	n	Pre	Change	CI	p
User Avg.	1103	69.71	-	6.16	-	1063	63.13	-	7.87	-
LIME	190	-	5.70	9.05	.197	179	-	5.25	10.59	.309
Anchor	199	-	0.86	10.48	.869	197	-	5.66	7.91	.140
Prototype	223	-	−2.64	9.59	.566	192	-	**9.53**	8.55	.032
DB	205	-	−0.92	11.87	.876	207	-	2.48	11.62	.667
Composite	286	-	−2.07	8.51	.618	288	-	7.36	9.38	.122

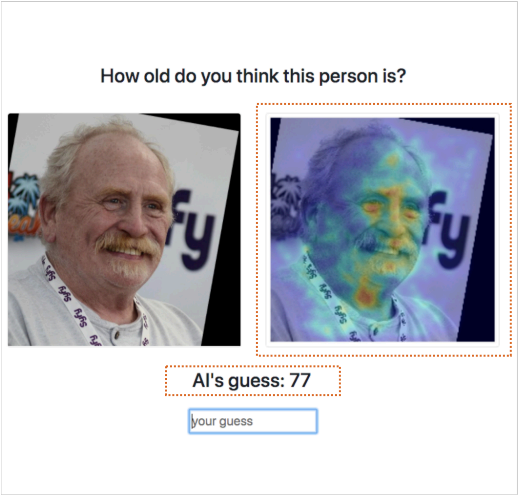
Table 2: Change in user accuracies after being given explanations of model behavior, relative to the baseline performance (Pre). Data is grouped by simulation test type. CI gives the 95% confidence interval, calculated by bootstrap using n user responses. We bold results that are significant at the $p < .05$ level. Prototype explanations improve counterfactual simulatability, while other methods do not definitively improve simulatability for one test.

Peter Hase, Mohit Bansal:Evaluating Explainable AI: Which Algorithmic Explanations Help Users Predict Model Behavior? ACL 2020: 5540-5552

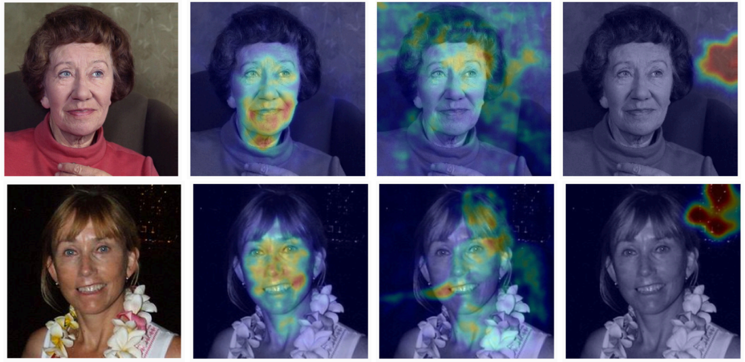
Evaluation (5) – … But No So Clear If Saliency Maps Are Always of Use



(a) Description of the model and guidelines for interpreting and using the explanations.



(b) Users are asked to guess a person's age.



(a) Original image (b) "Strong" (c) "Spurious" (d) "Random"

Treatment Arm	MAE
Control (Human Alone)	10.0 (9.4 - 10.5)
Model Alone	8.5 (8.3 - 8.7)
Prediction	8.4 (7.8 - 9.0)
Explain-strong	8.0 (7.5 - 8.5)
Explain-spurious	8.5 (8.0 - 9.1)
Explain-random	8.7 (8.1 - 9.2)
Delayed Prediction	8.5 (8.0 - 9.0)
Empathetic	8.0 (7.6 - 8.5)
Show Top-3 Range	8.0 (7.4 - 8.5)

- Faulty explanations did not significantly decrease trust in model predictions
- Most participants claimed that explanations appeared reasonable, even when they were obviously not focused on faces

Eric Chu, Deb Roy, Jacob Andreas: Are Visual Explanations Useful? A Case Study in Model-in-the-Loop Prediction. CoRR abs/2007.12248 (2020)

Evaluation (6) – Is Explanation Only for Debugging?

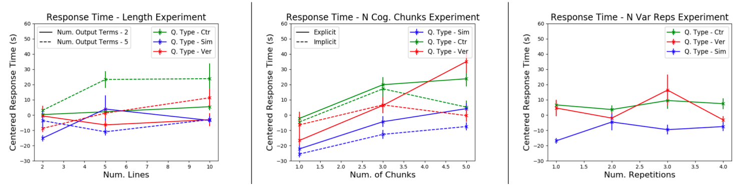
Domain	Model Purpose	Explainability Technique	Stakeholders	Evaluation Criteria
Finance	Loan Repayment	Feature Importance	Loan Officers	Completeness [34]
Insurance	Risk Assessment	Feature Importance	Risk Analysts	Completeness [34]
Content Moderation	Malicious Reviews	Feature Importance	Content Moderators	Completeness [34]
Finance	Cash Distribution	Feature Importance	ML Engineers	Sensitivity [69]
Facial Recognition	Smile Detection	Feature Importance	ML Engineers	Faithfulness [7]
Content Moderation	Sentiment Analysis	Feature Importance	QA ML Engineers	ℓ_2 norm
Healthcare	Medicare Access	Counterfactual Explanations	ML Engineers	normalized ℓ_1 norm
Content Moderation	Object Detection	Adversarial Perturbation	QA ML Engineers	ℓ_2 norm

Table 1: Summary of select deployed local explainability use cases

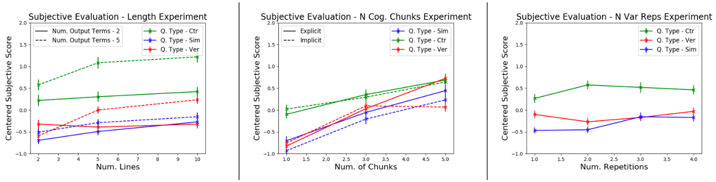
Umang Bhatt, Alice Xiang, Shubham Sharma, Adrian Weller, Ankur Taly, Yunhan Jia, Joydeep Ghosh, Ruchir Puri, José M. F. Moura, Peter Eckersley:Explainable machine learning in deployment. FAT* 2020: 648-657

Through Amazon Mechanical Turk
(900 subjects all together)

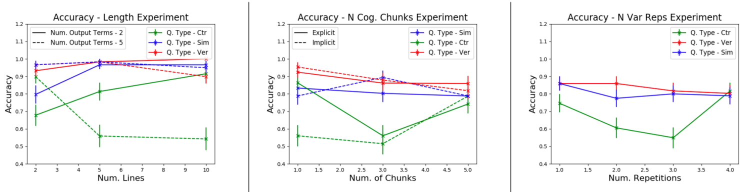
Response Time



Subjective Satisfaction



Accuracy



Isaac Lage, Emily Chen, Jeffrey He, Menaka Narayanan, Been Kim, Sam Gershman, Finale Doshi-Velez: An Evaluation of the Human-Interpretability of Explanation. CoRR abs/1902.00006 (2019)

Evaluation (7) - 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?

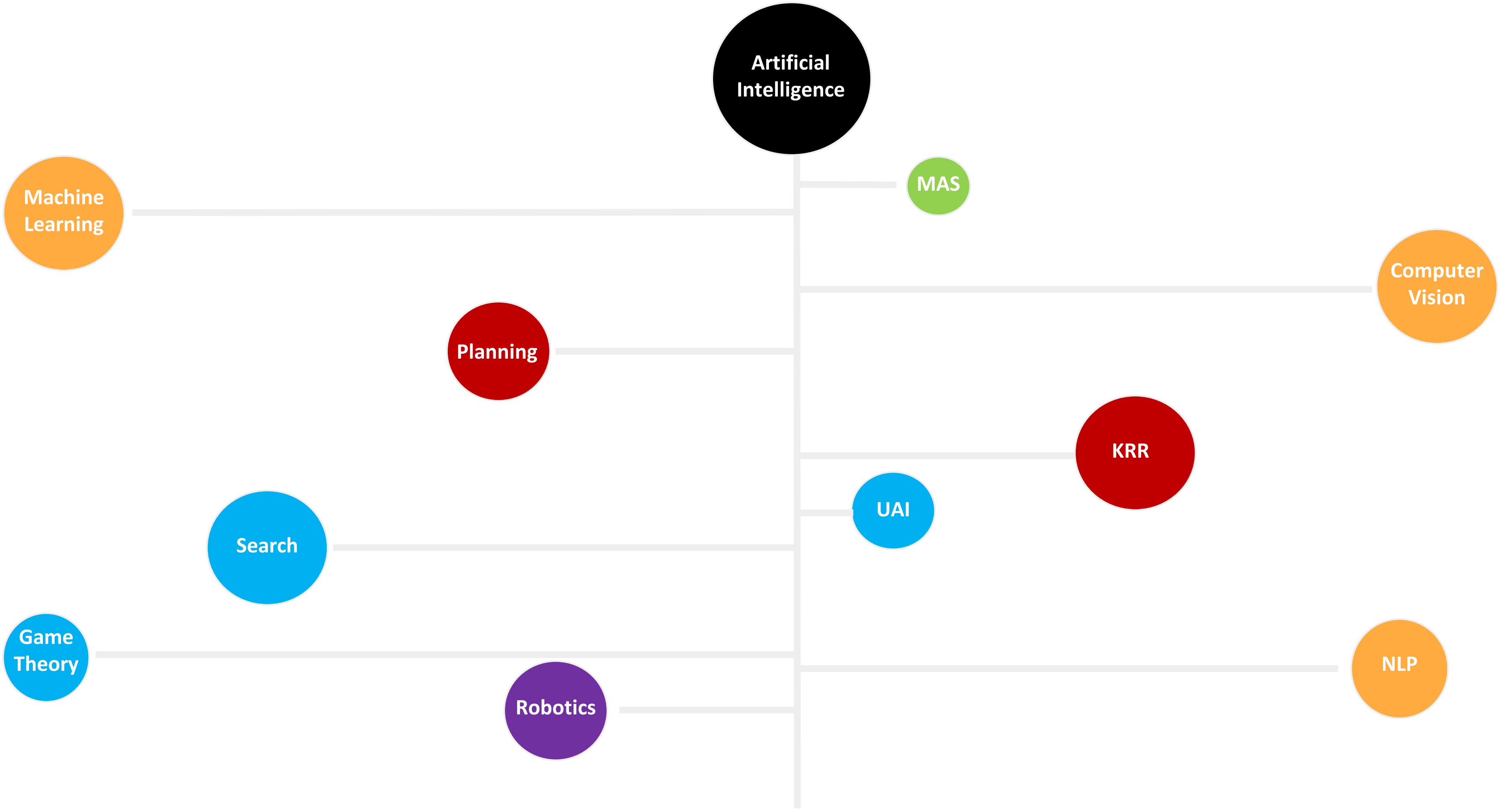
Source: Accenture Point of View. Understanding Machines: Explainable AI. Freddy Lecue, Dadong Wan

Part II

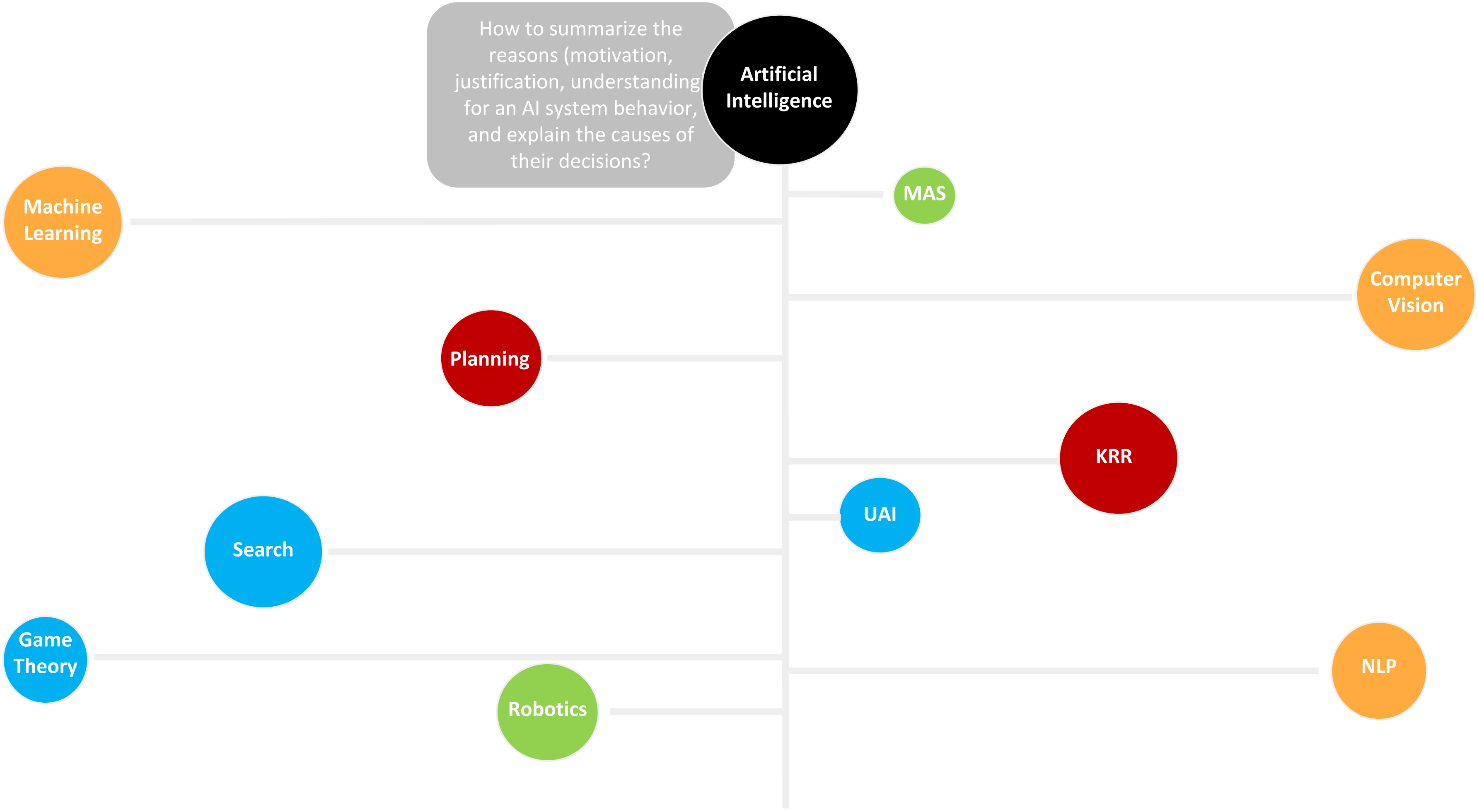
Explanation in AI (not only Machine Learning!)

Freddy Lécué: On the role of knowledge graphs in explainable AI. Semantic Web 11(1): 41-51 (2020)

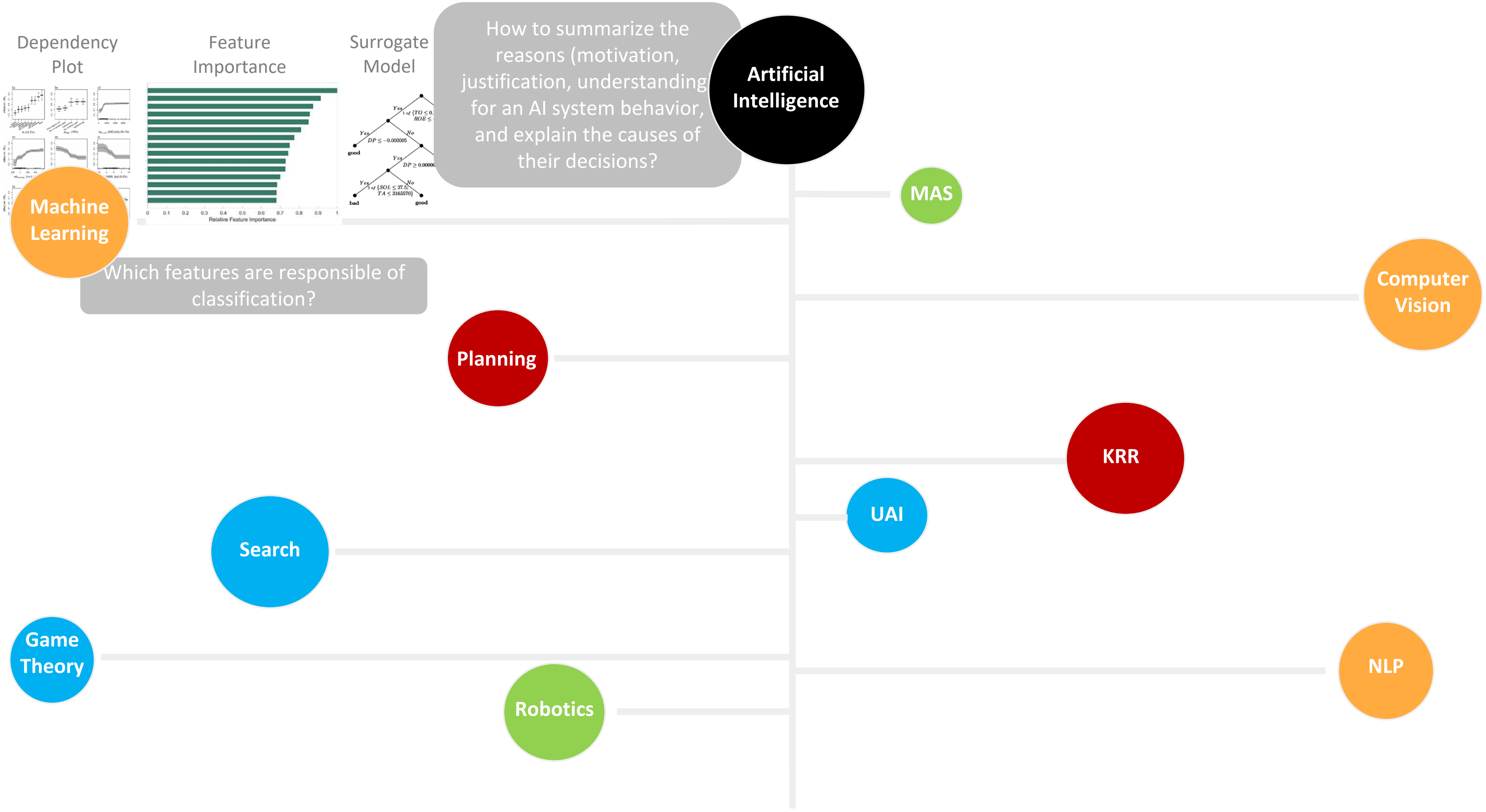
XAI: One Objective, Many 'AI's, Many Definitions, Many Approaches



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

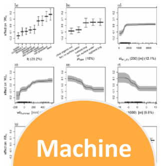


XAI: One Objective, Many ‘AI’s, Many Definitions, Many Approaches

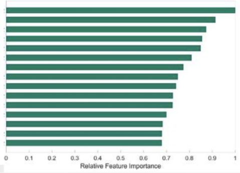


XAI: One Objective, Many 'AI's, Many Definitions, Many Approaches

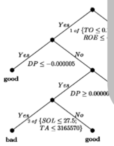
Dependency Plot



Feature Importance



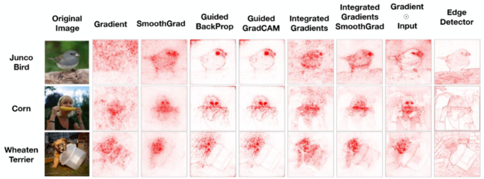
Surrogate Model



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

Artificial Intelligence

Saliency Map



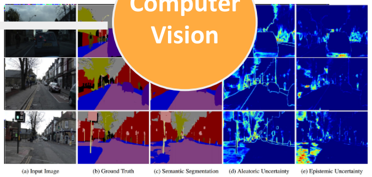
Which complex features are responsible of classification?

MAS

Machine Learning

Which features are responsible of classification?

Computer Vision



Uncertainty Map

Planning

KRR

UAI

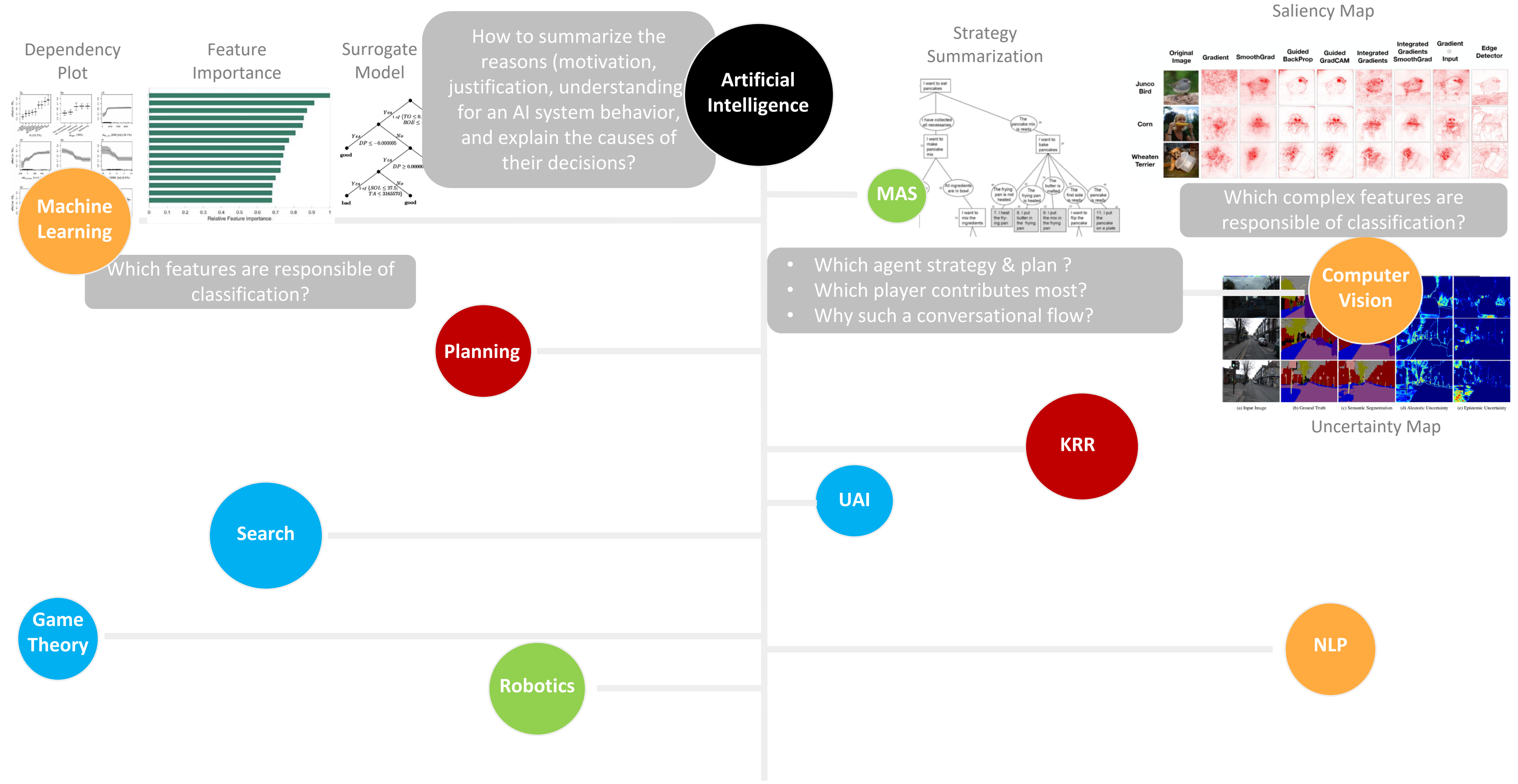
Search

Game Theory

NLP

Robotics

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

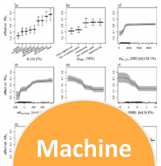


XAI: One Objective, Many 'AI's, Many Definitions, Many Approaches

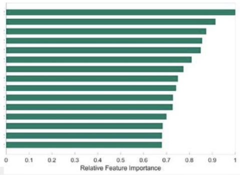
Artificial Intelligence

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

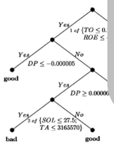
Dependency Plot



Feature Importance



Surrogate Model



Machine Learning

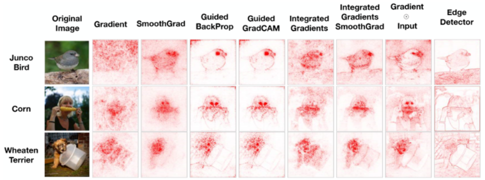
Which features are responsible of classification?

Strategy Summarization

MAS

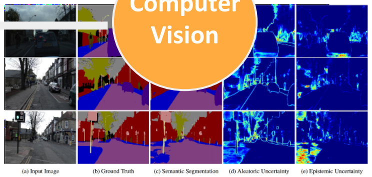
- Which agent strategy & plan ?
- Which player contributes most?
- Why such a conversational flow?

Saliency Map



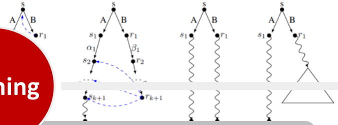
Which complex features are responsible of classification?

Computer Vision



Uncertainty Map

Plan Refinement



Planning

Which actions are responsible of a plan?

KRR

UAI

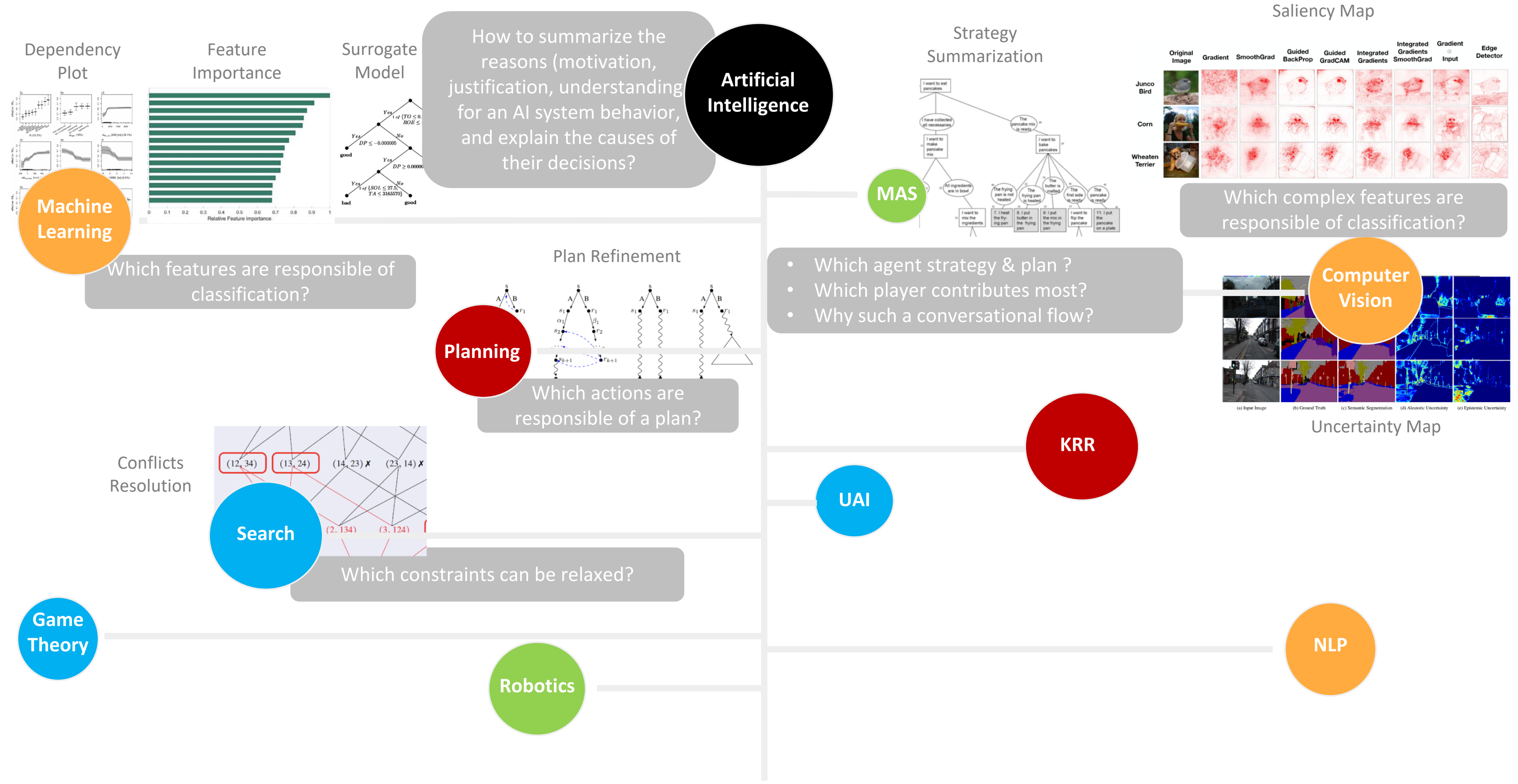
Search

Game Theory

Robotics

NLP

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



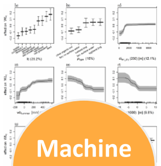
XAI: One Objective, Many ‘AI’s, Many Definitions, Many Approaches

Artificial Intelligence

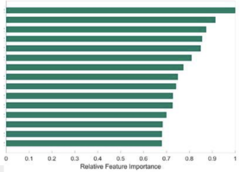
How to summarize the reasons (motivation, justification, understanding for an AI system behavior, and explain the causes of their decisions?

Machine Learning

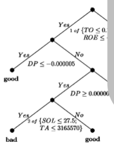
Dependency Plot



Feature Importance



Surrogate Model



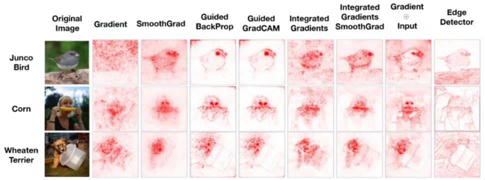
Which features are responsible of classification?

MAS

Strategy Summarization

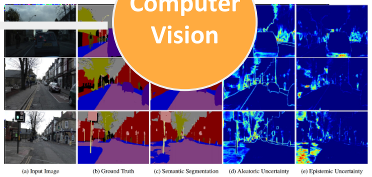
- Which agent strategy & plan ?
- Which player contributes most?
- Why such a conversational flow?

Saliency Map



Which complex features are responsible of classification?

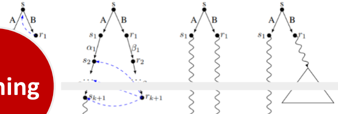
Computer Vision



Uncertainty Map

Planning

Plan Refinement



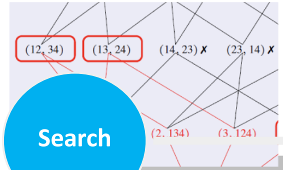
Which actions are responsible of a plan?

KRR

UAI

Search

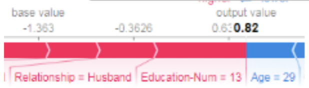
Conflicts Resolution



Which constraints can be relaxed?

Game Theory

Which combination of features is optimal?

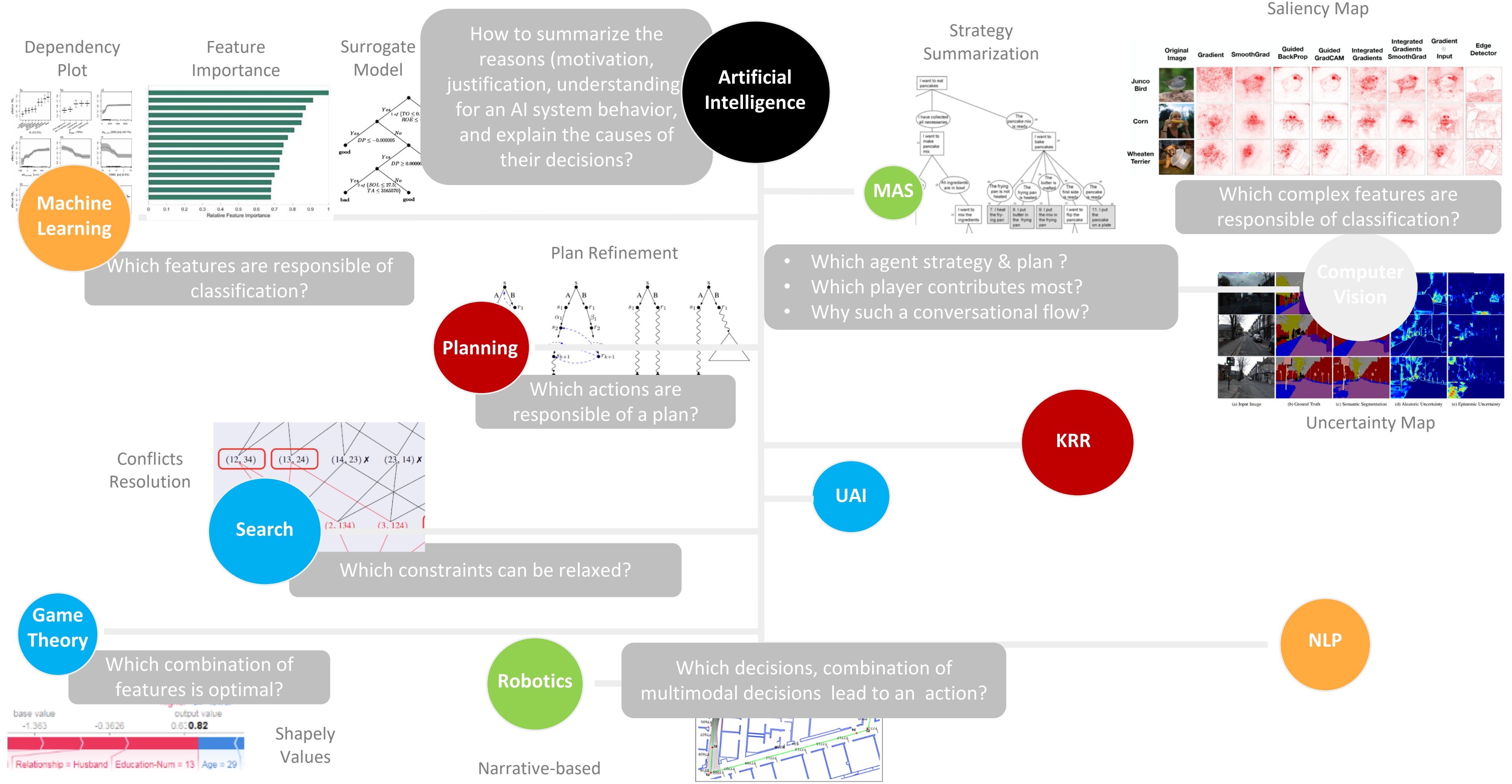


Shapely Values

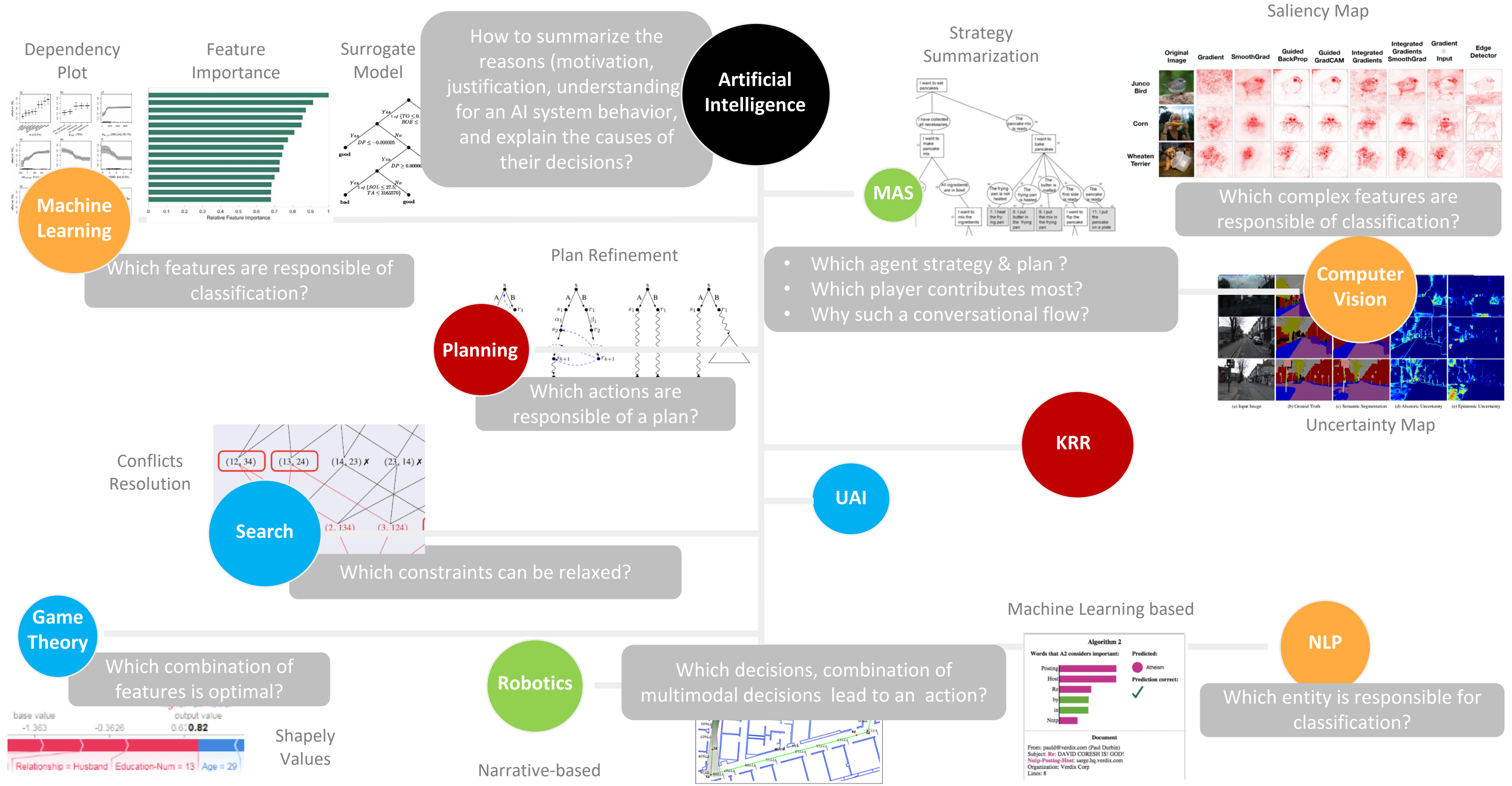
Robotics

NLP

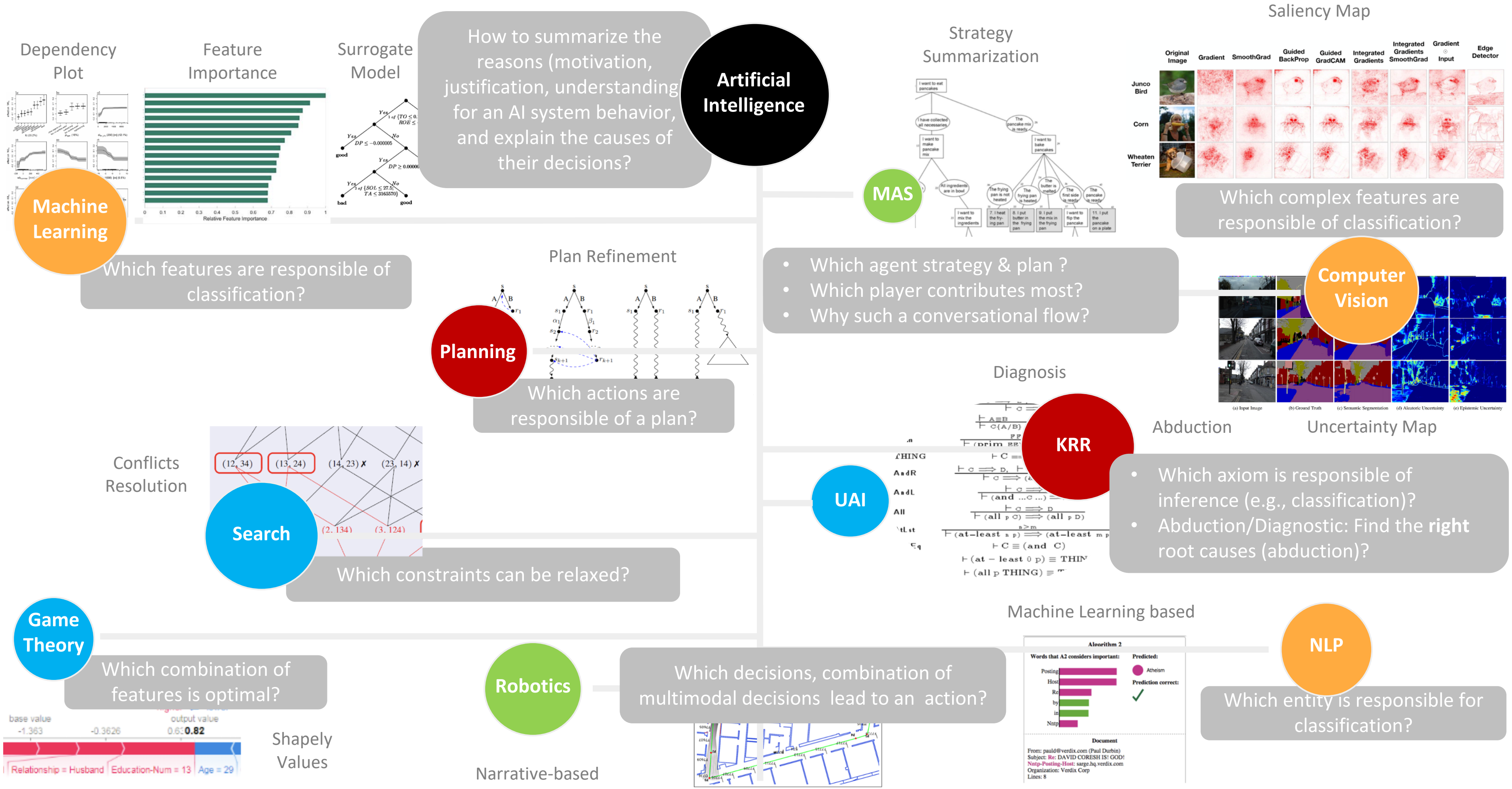
XAI: One Objective, Many 'AI's, Many Definitions, Many Approaches



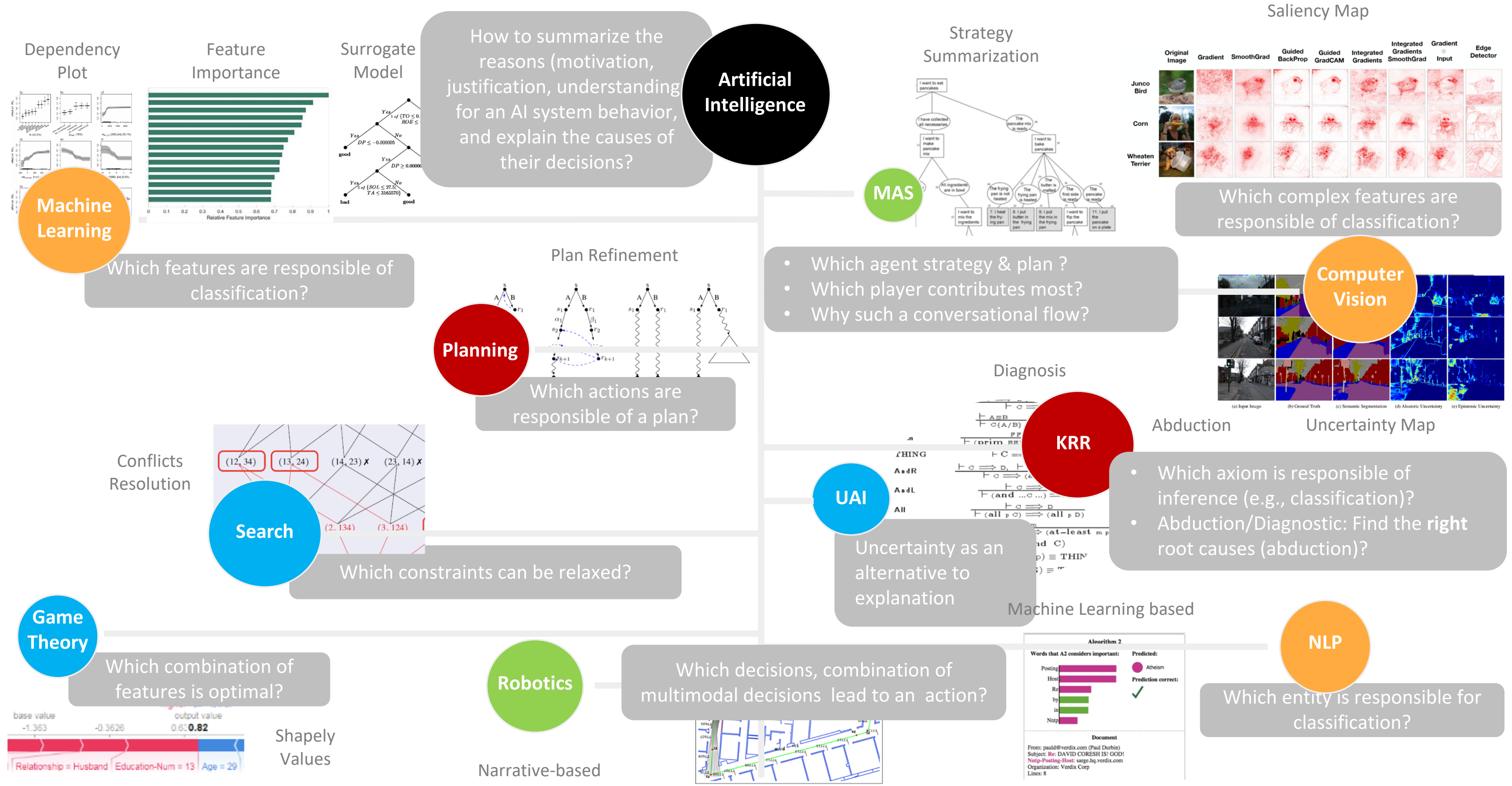
XAI: One Objective, Many 'AI's, Many Definitions, Many Approaches



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



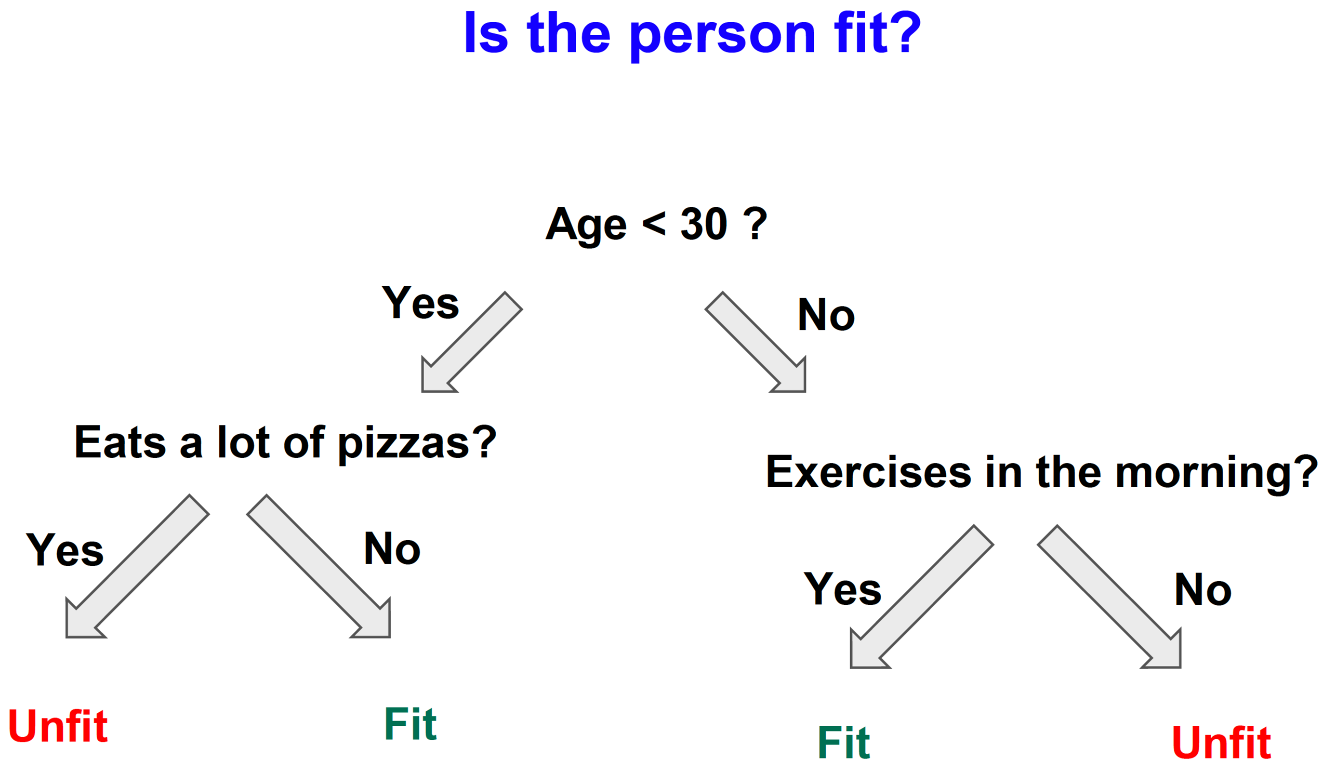
XAI: One Objective, Many 'AI's, Many Definitions, Many Approaches



Overview of explanation in Machine Learning (1)

- All except Artificial Neural Network

Interpretable Models:
- Decision Trees



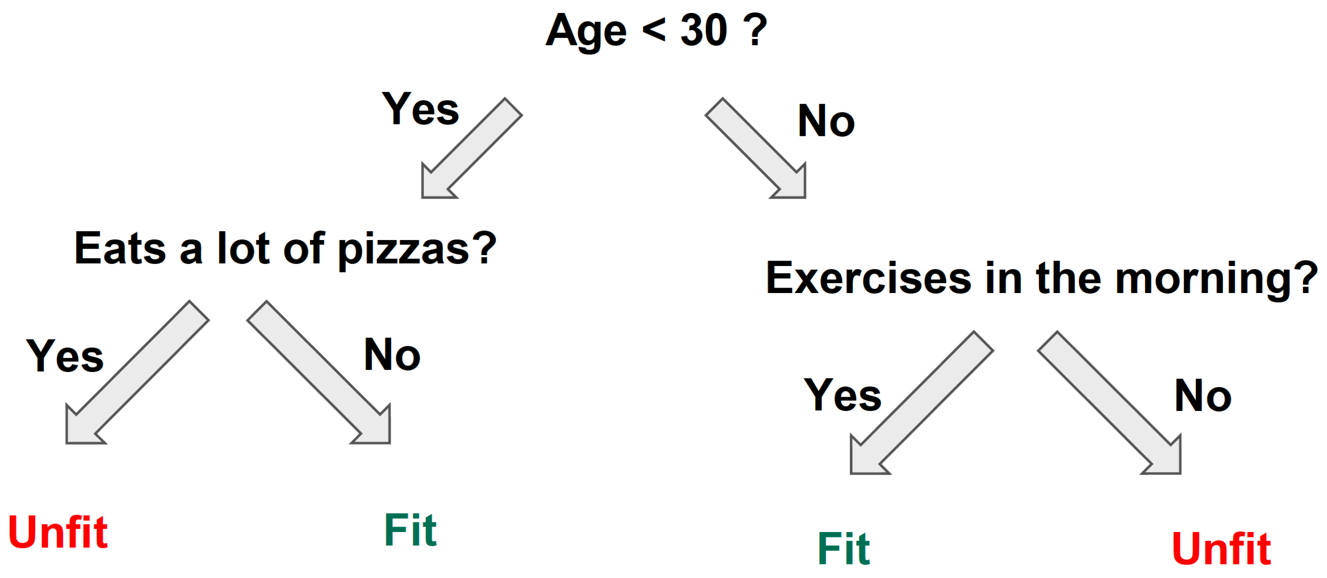
Overview of Explanation in Machine Learning (1)

- Many tools already available from early-days Machine Learning

Interpretable Models:

- Decision Trees

Is the person fit?



Overview of Explanation in Machine Learning (1)

- Many tools already available from early-days Machine Learning

Interpretable Models:

- Decision Trees, Lists

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

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

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

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

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

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

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

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

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

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

Else Diabetes

Overview of Explanation in Machine Learning (1)

- Many tools already available from early-days Machine Learning

Interpretable Models:

- Decision Trees, Lists and Sets and rules

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

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

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

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

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

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

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

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

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

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

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

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

Overview of Explanation in Machine Learning (1)

- Many tools already available from early-days Machine Learning

Interpretable Models:
- Decision Trees, Lists and Sets and rules
- GAMs,
- GLMs,

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

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

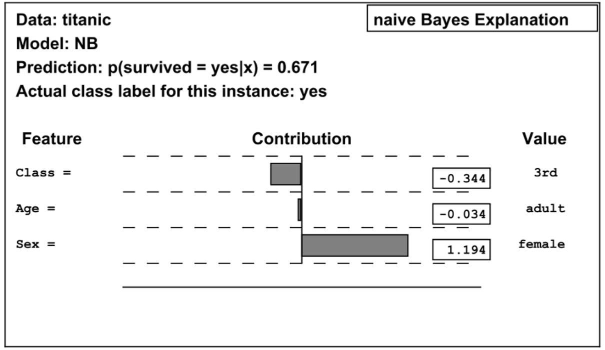
Accurate Intelligible Models with Pairwise Interactions. Lou, Caruana, Gehrke and Hooker. KDD 2013

Overview of Explanation in Machine Learning (1)

- Many tools already available from early-days Machine Learning

Interpretable Models:

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



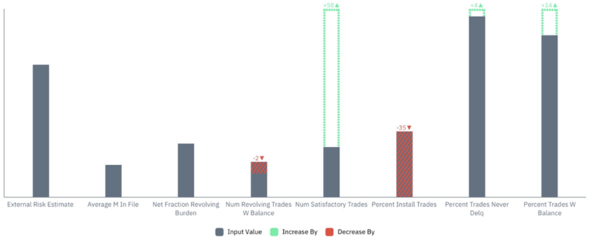
Naive Bayes model

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

Overview of Explanation in Machine Learning (1)

- Many tools already available from early-days Machine Learning

Interpretable Models:
- Decision Trees, Lists and Sets and rules
- GAMs,
- GLMs,
- Linear regression,
- Logistic regression,
- KNNs

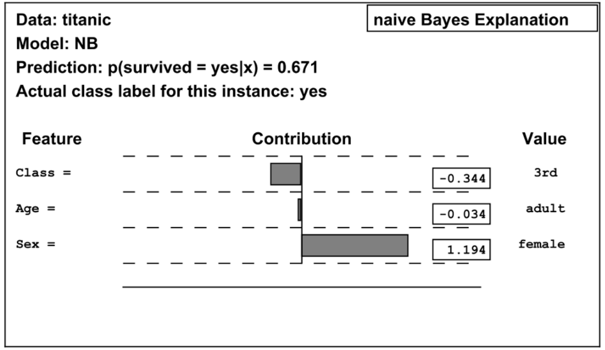


Counterfactual What-if

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

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

https://pair-code.github.io/what-if-tool/



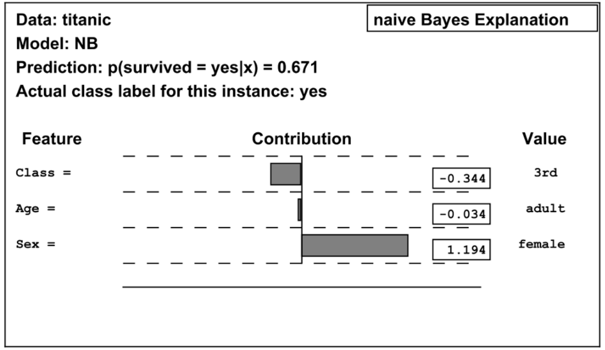
Naive Bayes model

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

Overview of Explanation in Machine Learning (1)

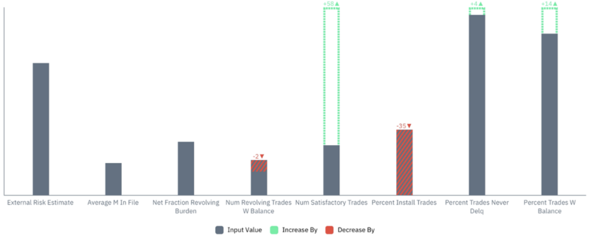
- Many tools already available from early-days Machine Learning

Interpretable Models:
- Decision Trees, Lists and Sets and rules
- GAMs,
- GLMs,
- Linear regression,
- Logistic regression,
- KNNs



Naive Bayes model

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

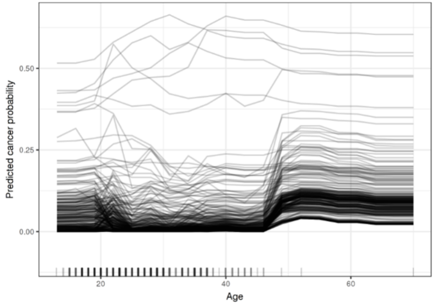


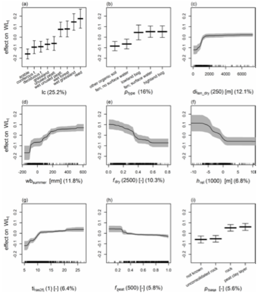
Counterfactual What-if

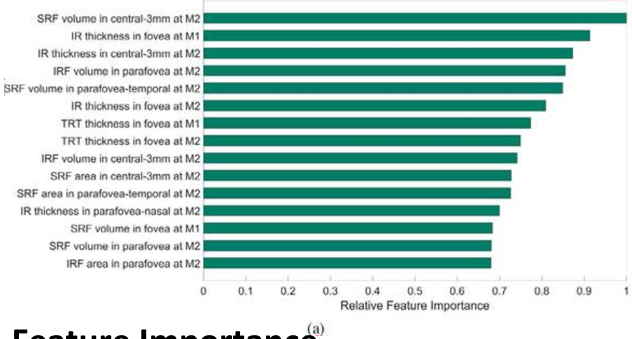
Brent D. Mittelstadt, Chris Russell, Sandra Wachter: Explaining Explanations in AI. FAT 2019: 279-288

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)

https://pair-code.github.io/what-if-tool/



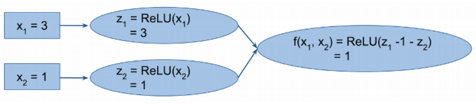




- **Feature Importance**
- **Partial Dependence Plot**
- **Individual Conditional Expectation**
- **Sensitivity Analysis**

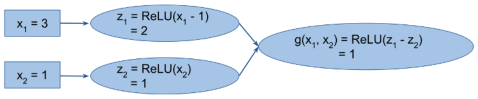
Overview of Explanation in Machine Learning (2)

- Focus: 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	$x_1 = 1.5,\ x_2 = -0.5$



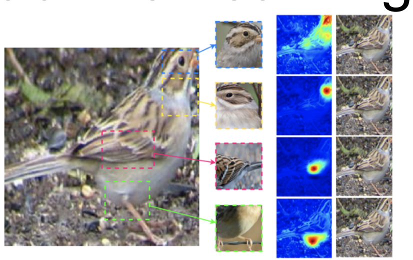
Network $g(x_1, x_2)$
Attributions at $x_1 = 3, x_2 = 1$

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

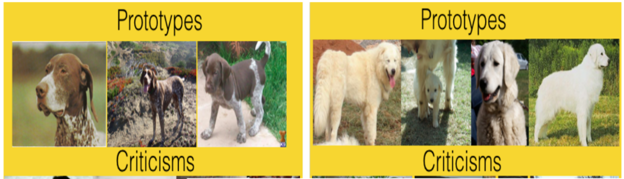
Attribution for Deep Network (Integrated gradient-based)

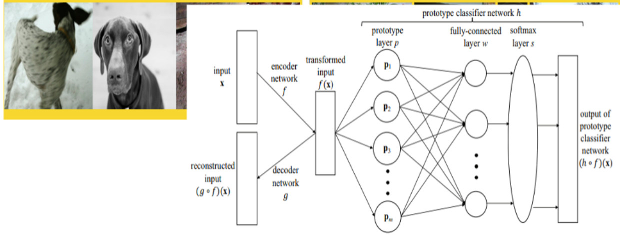
Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In ICML, pp. 3319–3328, 2017.

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)

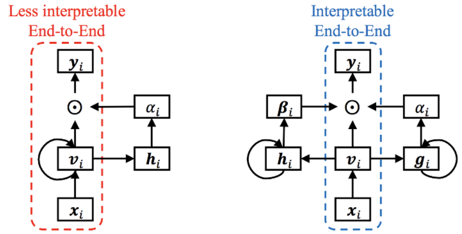




Example-based / Prototype

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

Been Kim, Oluwasanmi Koyejo, Rajiv Khanna:Examples are not enough, learn to criticize! Criticism for Interpretability. NIPS 2016: 2280-2288

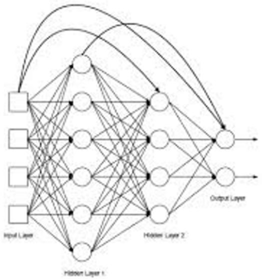


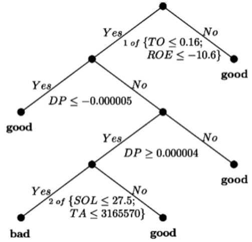
(a) Standard attention model (b) RETAIN model

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

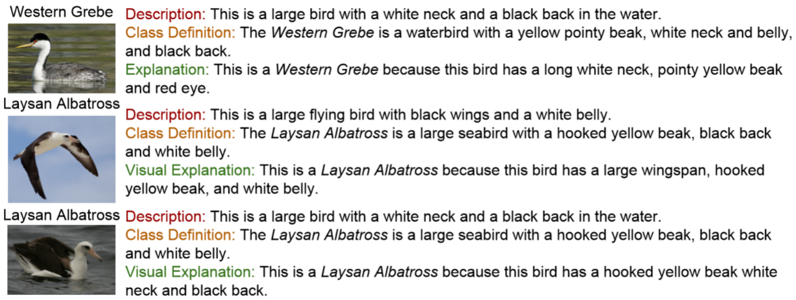
Mark Craven, Jude W. Shavlik: Extracting Tree-Structured Representations of Trained Networks. NIPS 1995: 24-30

Overview of Explanation in Machine Learning (3)

- Focus: Artificial Neural Network

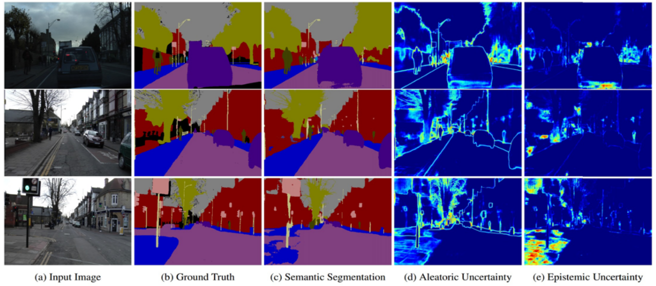
Interpretable Units

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



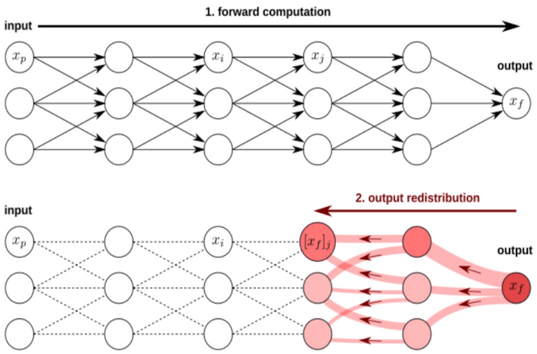
Visual Explanation

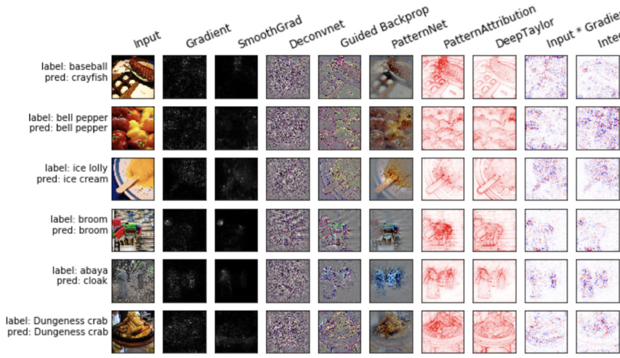
Lisa Anne Hendricks, Zeynep Akata, Marcus Rohrbach, Jeff Donahue, Bernt Schiele, Trevor Darrell: Generating Visual Explanations. ECCV (4) 2016: 3-19



Uncertainty Map

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



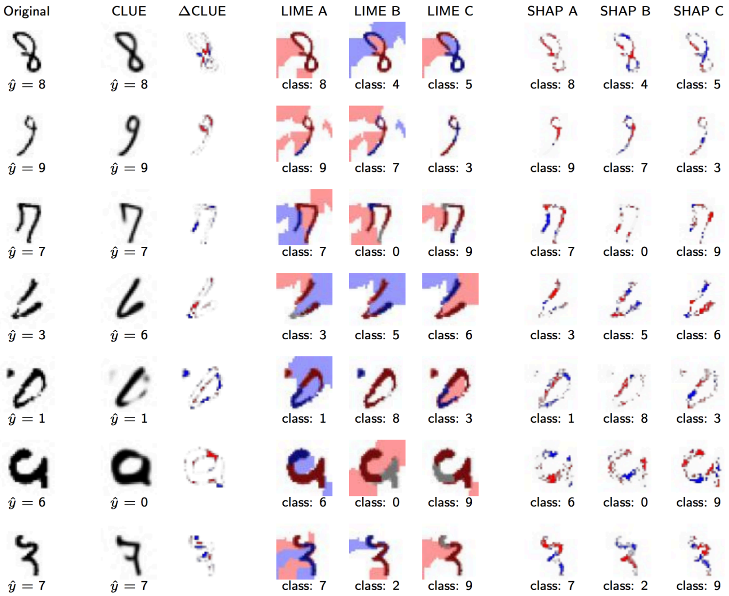


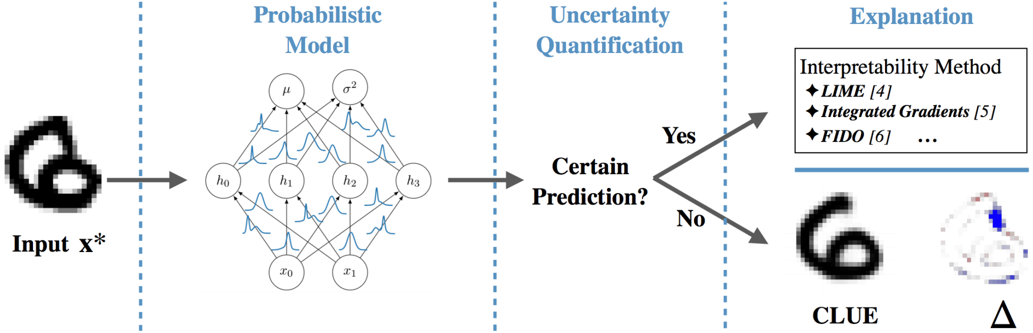
Saliency Map / Features Attribution-based

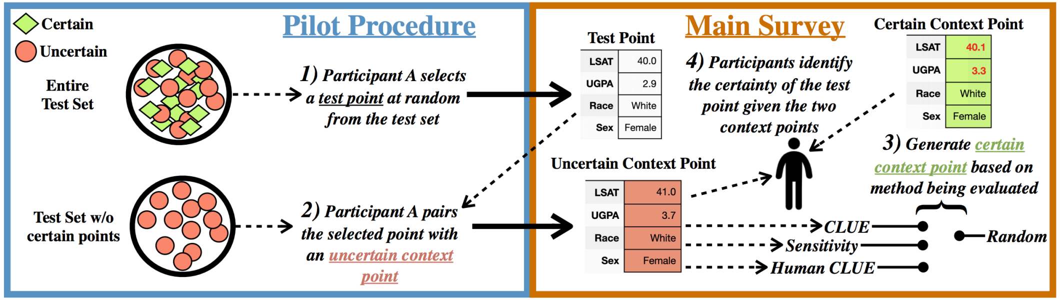
Julius Adebayo, Justin Gilmer, Michael Muelly, Ian J. Goodfellow, Moritz Hardt, Been Kim: Sanity Checks for Saliency Maps. NeurIPS 2018: 9525-9536

Overview of Explanation in Machine Learning (4)

- Focus: Artificial Neural Network







Explaining Uncertainty - Beyond Interpretation of Prediction

Javier Antoran, Umang Bhatt, Tameem Adel, Adrian Weller, José Miguel Hernández-Lobato: Getting a clue: a method for explaining uncertainty estimates. ICLR 2021

Overview of Explanation in Machine Learning (5)

- Towards more semantic interpretation

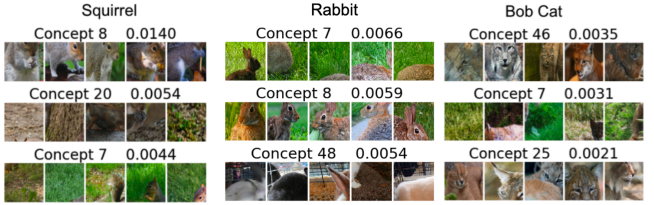
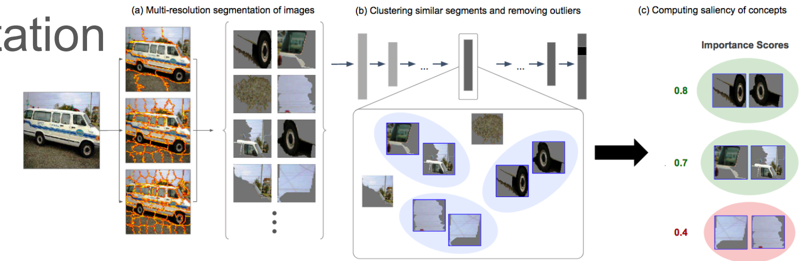
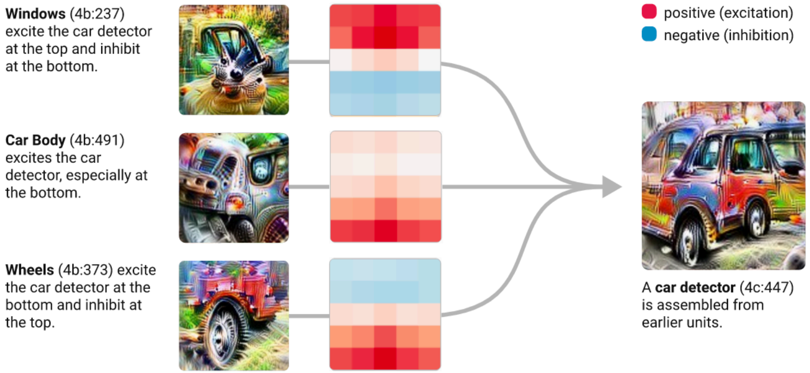


Figure 3: Concept examples with the samples that are the nearest to concept vectors in the activation space in AwA. The per-class ConceptSHAP score is listed above the images.

ConceptSHAP Chih-Kuan Yeh, Been Kim, Sercan Ömer Arik, Chun-Liang Li, Tomas Pfister, Pradeep Ravikumar:On Completeness-aware Concept-Based Explanations in Deep Neural Networks. NeurIPS 2020



ACE Amirata Ghorbani, James Wexler, James Y. Zou, Been Kim:Towards Automatic Concept-based Explanations. NeurIPS 2019: 9273-9282



Circuits in CNNs

https://distill.pub/2020/circuits/zoom-in/

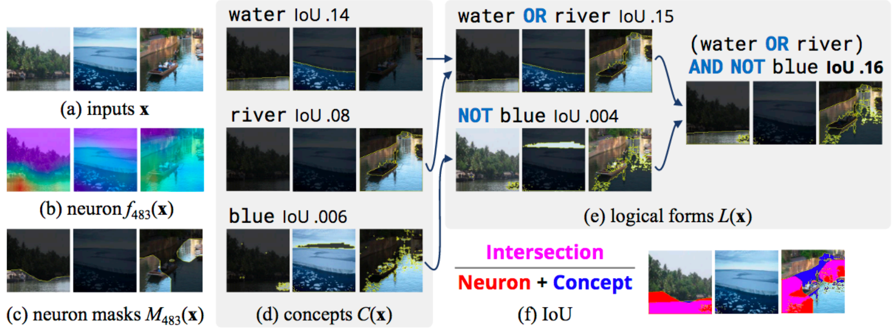


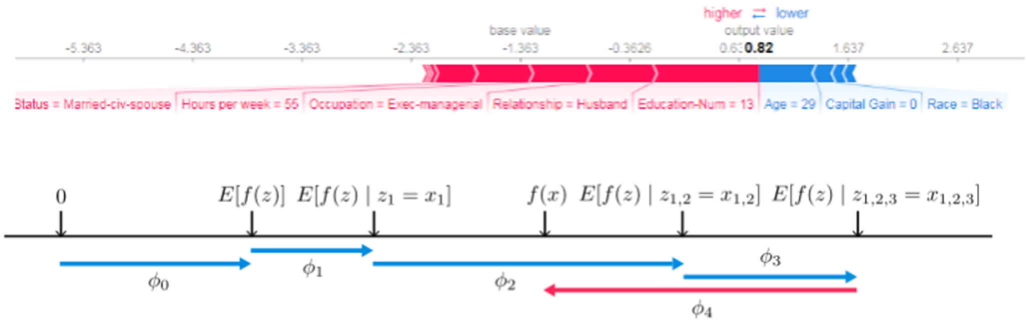
Figure 1: Given a set of inputs (a) and scalar neuron activations (b) converted into binary masks (c), we generate an explanation via beam search, starting with an inventory of primitive concepts (d), then incrementally building up more complex logical forms (e). We attempt to maximize the IoU score of an explanation (f); depicted is the IoU of $M_{483}(\mathbf{x})$ and `(water OR river) AND NOT blue`.

Compositional Explanations

Jesse Mu, Jacob Andreas:Compositional Explanations of Neurons. NeurIPS 2020

Overview of Explanation in Different AI Fields (1)

- Game Theory

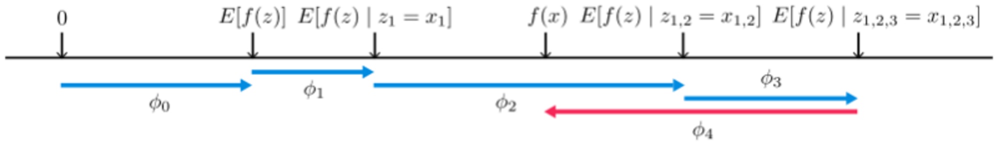


Shapley Additive Explanation

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

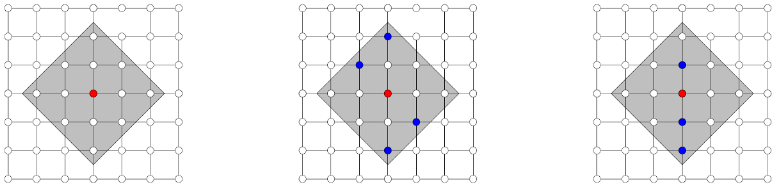
Overview of Explanation in Different AI Fields (1)

- Game Theory



Shapley Additive Explanation

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

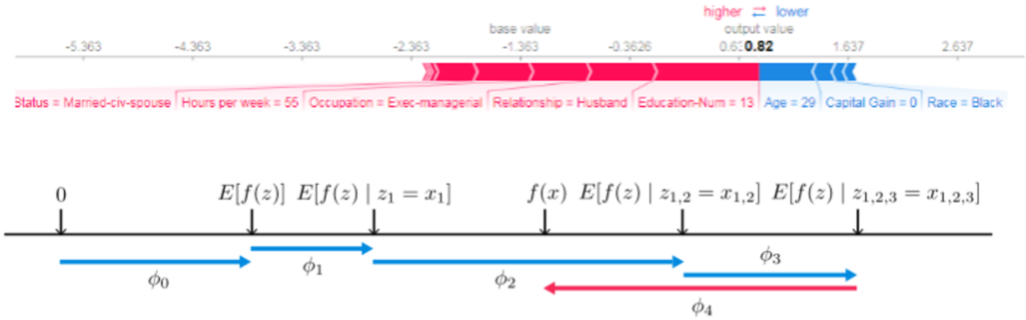


L-Shapley and C-Shapley (with graph structure)

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

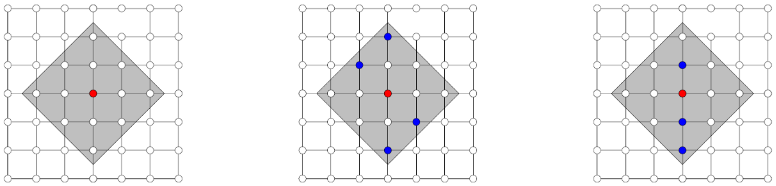
Overview of Explanation in Different AI Fields (1)

- Game Theory



Shapley Additive Explanation

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



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

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

~ instancewise feature importance (causal influence)

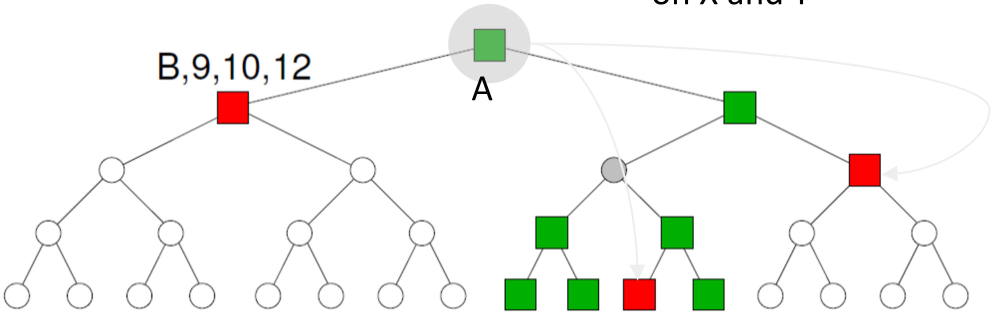
Erik Štrumbelj and Igor Kononenko. An efficient explanation of individual classifications using game theory. Journal of Machine Learning Research, 11:1–18, 2010.

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

Overview of Explanation in Different AI Fields (2)

- Search and Constraints 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

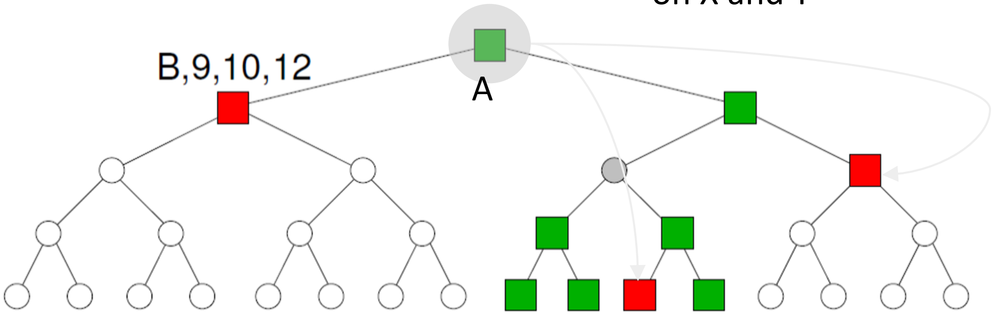
Robustness Computation

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 (2)

- Search and Constraints Satisfaction

If A+1 then NEW Conflicts on X and Y

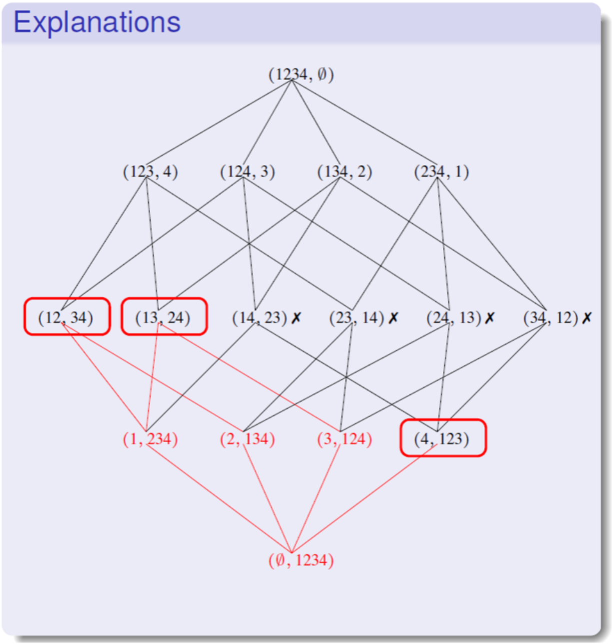


Conflicts resolution

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

Robustness Computation

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



Constraints relaxation

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

Overview of Explanation in Different AI Fields (3)

- Knowledge Representation and Reasoning

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

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

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

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

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

Overview of Explanation in Different AI Fields (3)

- Knowledge Representation and Reasoning

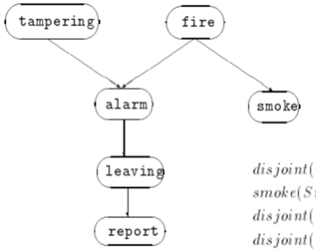
Ref	$\vdash C \Longrightarrow C$
Trans	$\dfrac{\vdash C \Longrightarrow D,\ \vdash D \Longrightarrow E}{\vdash C \Longrightarrow E}$
Eq	$\dfrac{\vdash A \equiv B \qquad \vdash C \Longrightarrow D}{\vdash C\{A/B\} \Longrightarrow D\{A/B\}}$
Prim	$\dfrac{FF \subset EE}{\vdash (\text{prim } EE) \Longrightarrow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

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

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

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

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



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

$disjoint([fire(yes) : 0.01, fire(no) : 0.99]).$

$smoke(Sm) \leftarrow fire(Fi) \wedge c\_smoke(Sm, Fi).$

$disjoint([c\_smoke(yes, yes) : 0.9, c\_smoke(no, yes) : 0.1]).$

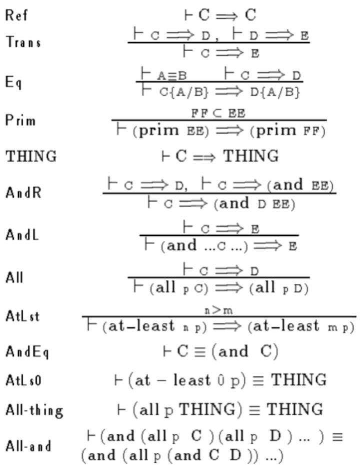
$disjoint([c\_smoke(yes, no) : 0.01, c\_smoke(no, no) : 0.99]).$

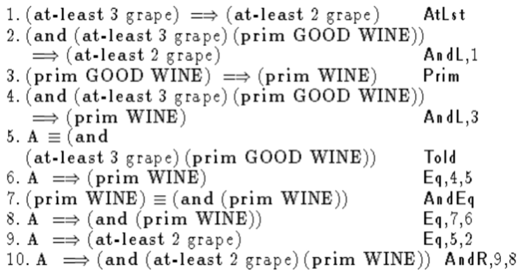
Abduction Reasoning (in Bayesian Network)

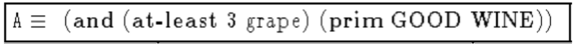
David Poole: Probabilistic Horn Abduction and Bayesian Networks. Artif. Intell. 64(1): 81-129 (1993)

Overview of Explanation in Different AI Fields (3)

- Knowledge Representation and Reasoning

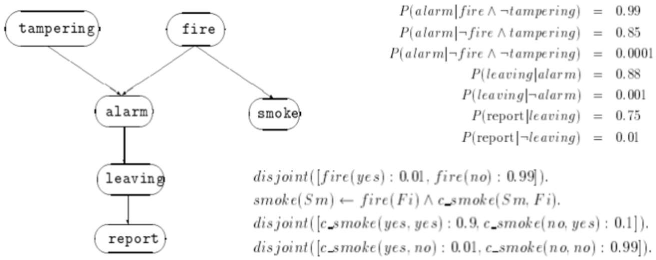






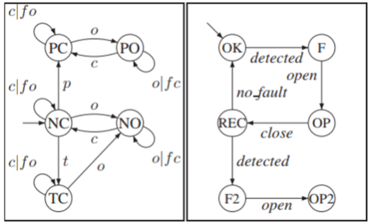
Explaining Reasoning (through Justification) e.g., Subsumption

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



Abduction Reasoning (in Bayesian Network)

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

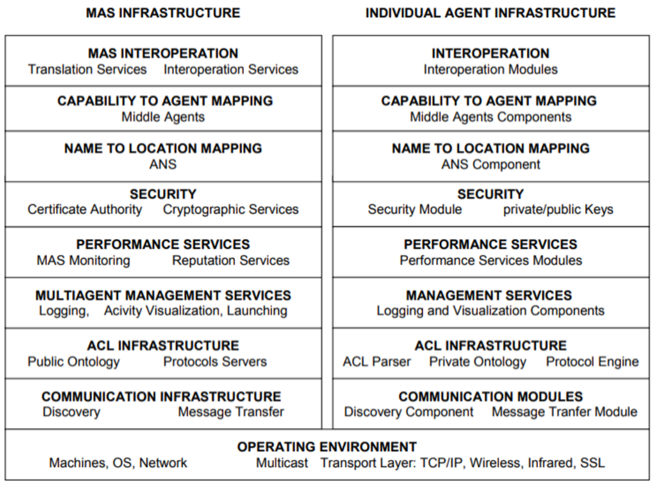


Diagnosis Inference

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

Overview of Explanation in Different AI Fields (4)

- Multi-Agents Systems

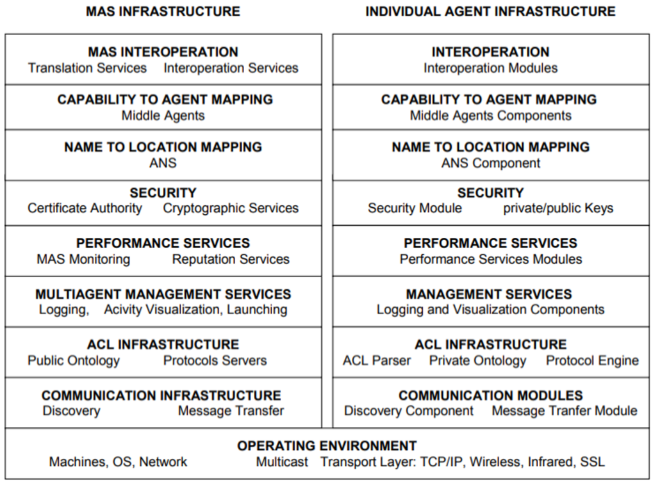


Explanation of Agent Conflicts & Harmful Interactions

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

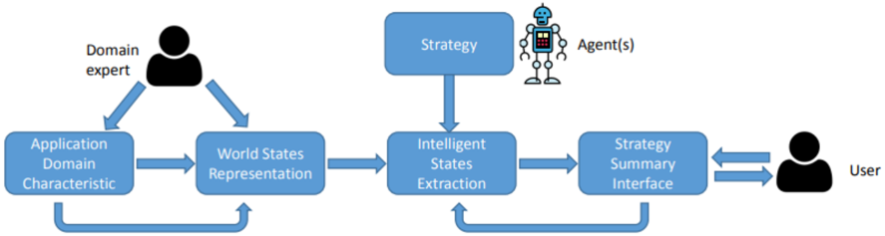
Overview of Explanation in Different AI Fields (4)

- Multi-Agents Systems



Explanation of Agent Conflicts & Harmful Interactions

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



Agent Strategy Summarization

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

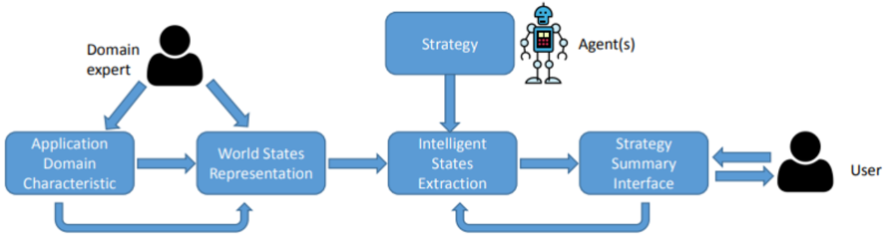
Overview of Explanation in Different AI Fields (4)

- Multi-Agents Systems

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

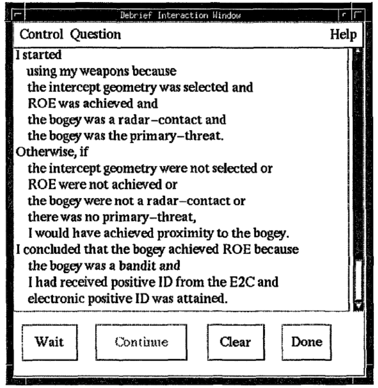
Explanation of Agent Conflicts & Harmful Interactions

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



Agent Strategy Summarization

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



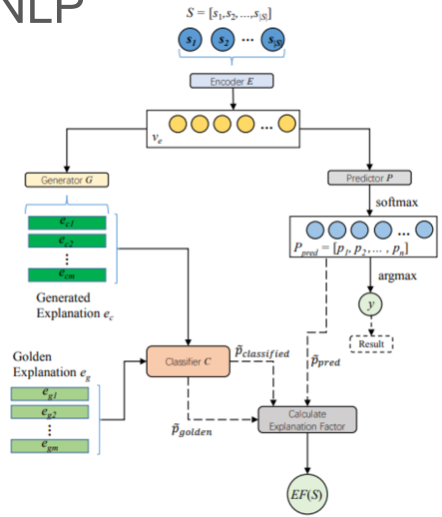
Explainable Agents

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

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

Overview of Explanation in Different AI Fields (5)

- NLP



Fine-grained explanations are in the form of:

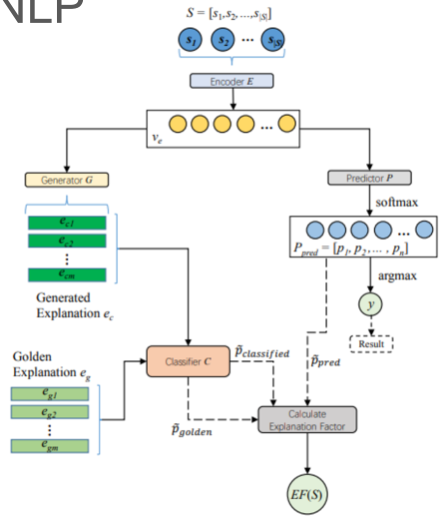
- texts in a real-world dataset;
- Numerical scores

Explainable NLP

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

Overview of Explanation in Different AI Fields (5)

- NLP

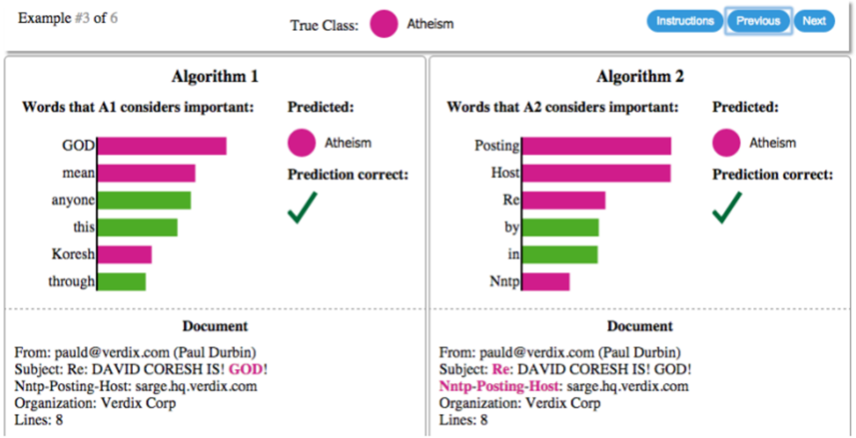


Fine-grained explanations are in the form of:

- 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)

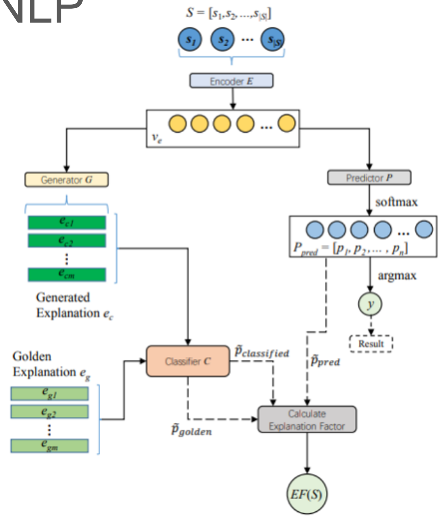


LIME for NLP

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 (5)

- NLP

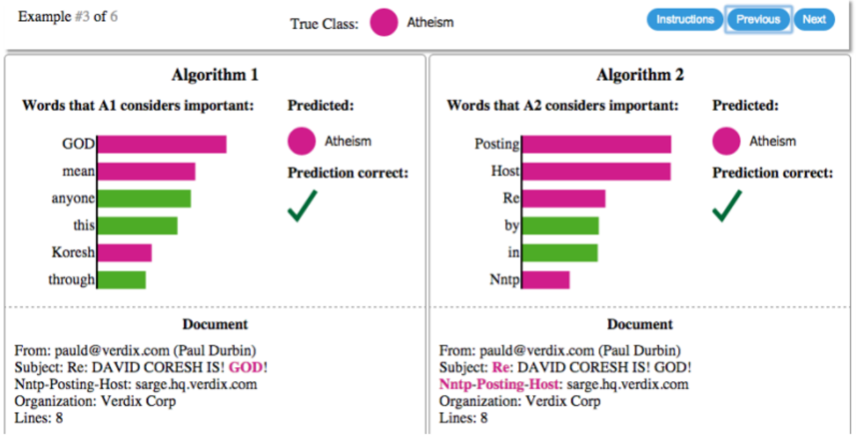


Fine-grained explanations are in the form of:

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

Explainable NLP

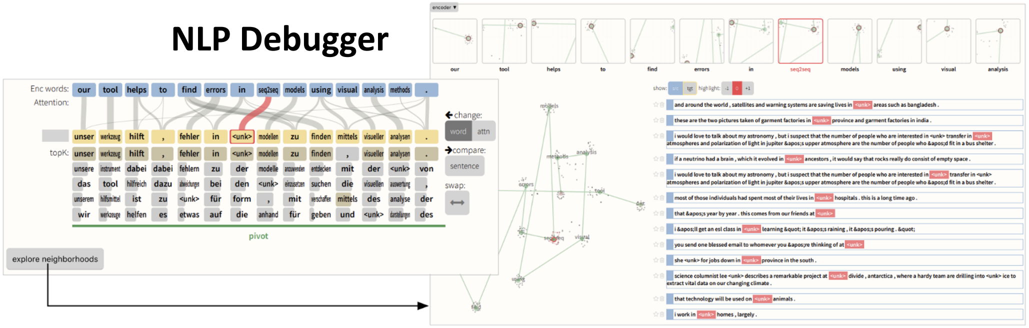
Hui Liu, Qingyu Yin, William Yang Wang: Towards Explainable NLP: A Gene
Explanation Framework for Text Classification. CoRR abs/1811.00196 (201



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

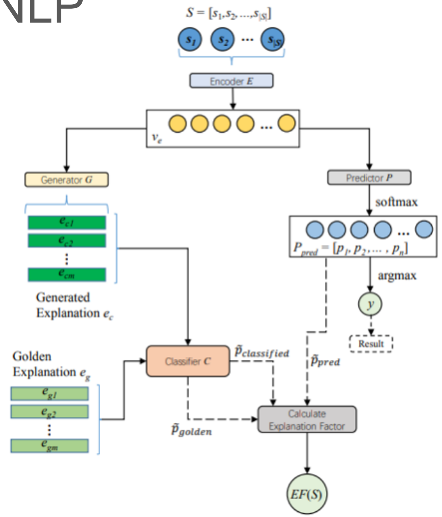
NLP Debugger



Hendrik Strobelt, Sebastian Gehrmann, 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 (5)

- NLP

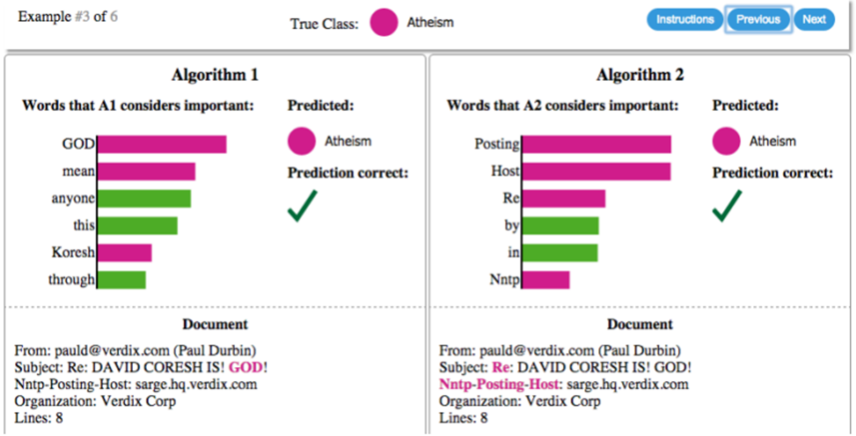


Fine-grained explanations are in the form of:

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

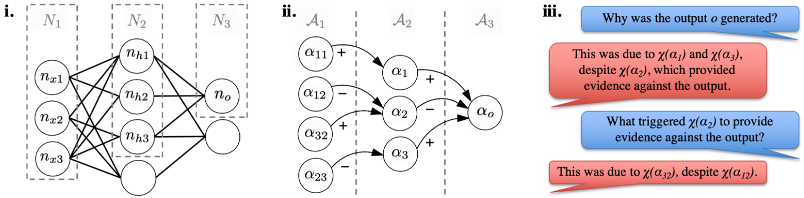
Explainable NLP

Hui Liu, Qingyu Yin, William Yang Wang: Towards Explainable NLP: A Gene
Explanation Framework for Text Classification. CoRR abs/1811.00196 (201



LIME for NLP

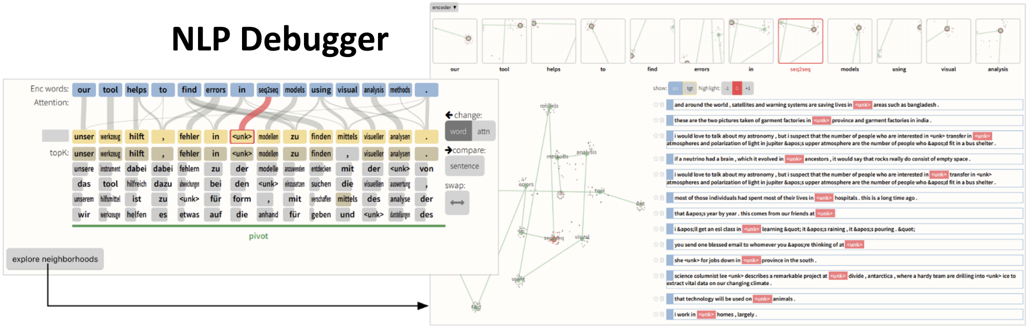
Marco Túlio Ribeiro, Sameer Singh, Carlos Guestrin: "Why Should I Trust You?": Explaining the Predictions of Any Classifier. KDD 2016: 1135-1144



Argumentation & Explanation

Emanuele Albini, Piyawat Lertvittayakumjorn, Antonio Rago, Francesca Toni:DAX: Deep Argumentative eXplanation for Neural Networks. CoRR abs/2012.05766 (2020)

NLP Debugger



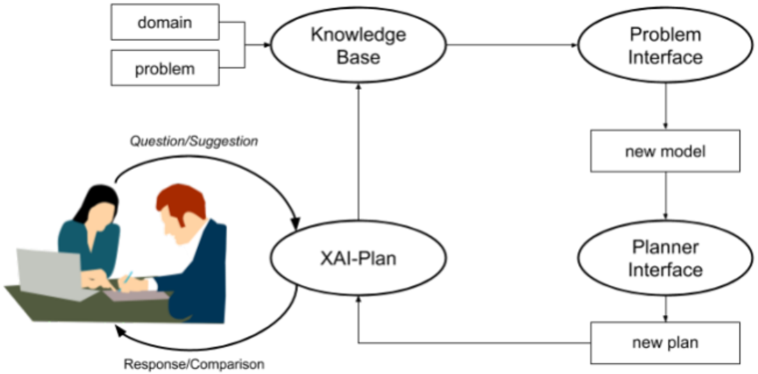
Hendrik Strobelt, Sebastian Gehrmann, Michael Behrisch, Adam Perer, Hanspeter Pfister, Alexander M. Rush: Seq2seq-Vis: A Visual Debugging Tool for Sequence-to-Sequence Models. IEEE Trans. Vis. Comput. Graph. 25(1): 353-363 (2019)

Overview of Explanation in Different AI Fields (6)

- Planning and Scheduling

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

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



XAI Plan

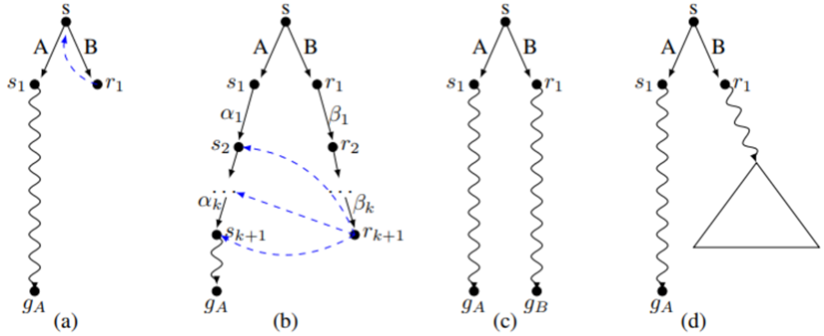
Rita Borgo, Michael Cashmore, Daniele Magazzeni: Towards Providing Explanations for AI Planner Decisions. CoRR abs/1810.06338 (2018)

Overview of Explanation in Different AI Fields (6)

- Planning and Scheduling

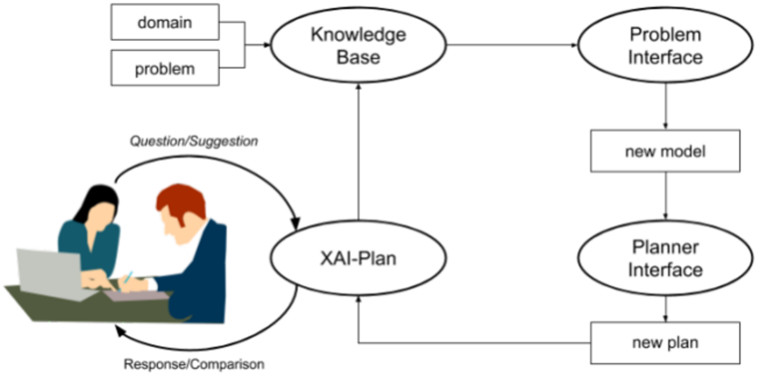
Explanation Type	R1	R2	R3	R4
Plan Patch Explanation / VAL	✗	✓	✗	✓
Model Patch Explanation	✓	✗	✓	✓
Minimally Complete Explanation	✓	✓	✗	?
Minimally Monotonic Explanation	✓	✓	✓	?
(Approximate) Minimally Complete Explanation	✗	✓	✗	✓

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



Human-in-the-loop Planning

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



XAI Plan

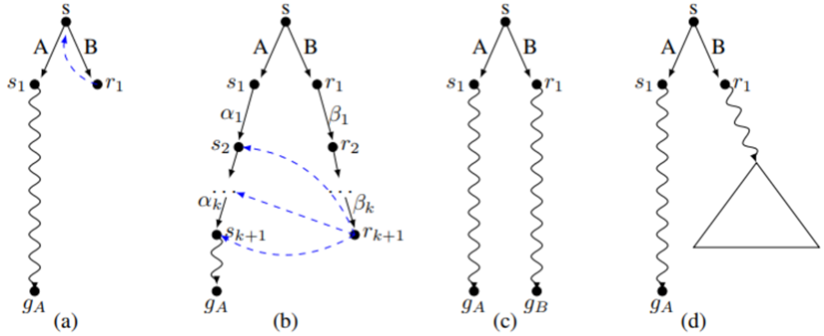
Rita Borgo, Michael Cashmore, Daniele Magazzeni: Towards Providing Explanations for AI Planner Decisions. CoRR abs/1810.06338 (2018)

Overview of Explanation in Different AI Fields (6)

- Planning and Scheduling

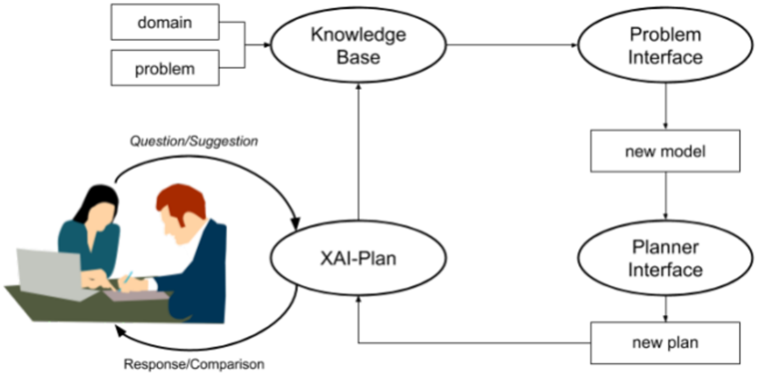
Explanation Type	R1	R2	R3	R4
Plan Patch Explanation / VAL	✗	✓	✗	✓
Model Patch Explanation	✓	✗	✓	✓
Minimally Complete Explanation	✓	✓	✗	?
Minimally Monotonic Explanation	✓	✓	✓	?
(Approximate) Minimally Complete Explanation	✗	✓	✗	✓

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



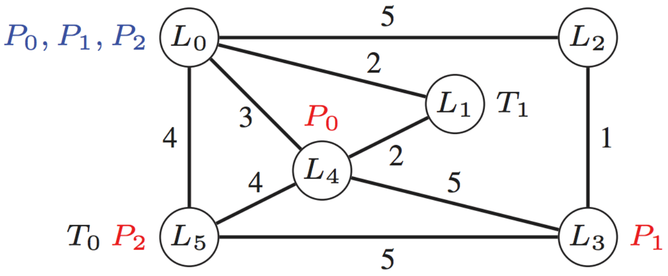
Human-in-the-loop Planning

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



XAI Plan

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



Explanation of the Space of Possible Plans

Rebecca Eifler, Michael Cashmore, Jörg Hoffmann, Daniele Magazzeni, Marcel Steinmetz: A New Approach to Plan-Space Explanation: Analyzing Plan-Property Dependencies in Oversubscription Planning. AAAI 2020: 9818-9826

Overview of Explanation in Different AI Fields (7)

- Robotics

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

Narration of Autonomous Robot Experience

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

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

Overview of Explanation in Different AI Fields (7)

- Robotics

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

Narration of Autonomous Robot Experience

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

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

Robot: I have decided to turn left.

Human: Why did you do that?

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

Human: How confident are you in this decision?

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

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

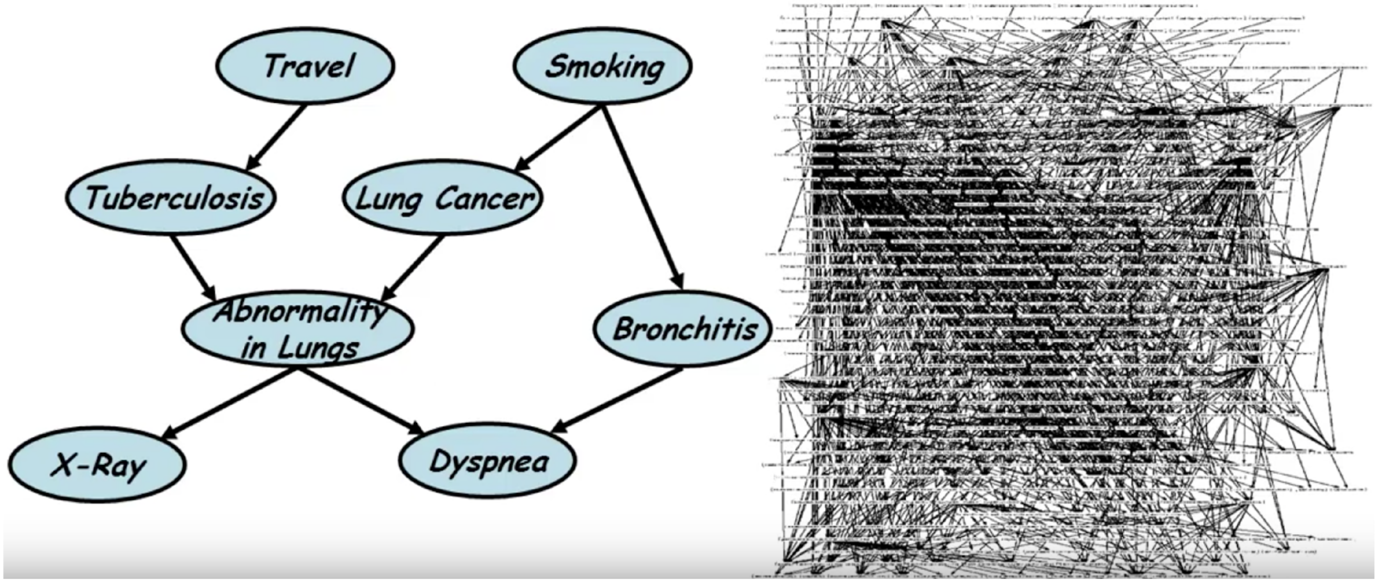
Robot: Here is the histogram of all training examples that reached this leaf. 80% of examples where this area was above 20 cm predicted the appropriate action to be "drive forward".

From Decision Tree to human-friendly information

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

Overview of Explanation in Different AI Fields (8)

- Reasoning under Uncertainty



Probabilistic Graphical Models

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

Part III

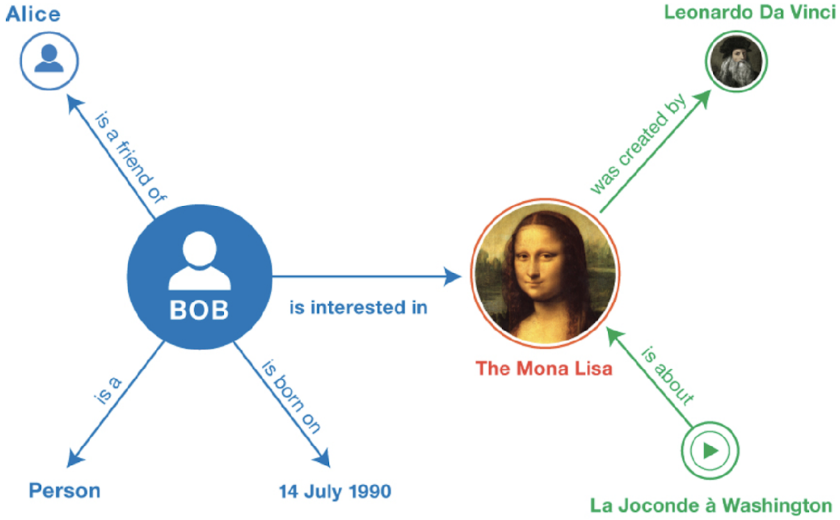
On The Role of Knowledge Graphs in Explainable Machine Learning

Freddy Lécué: On the role of knowledge graphs in explainable AI. Semantic Web 11(1): 41-51 (2020)

Knowledge Graph (1)

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

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



Freddy Lécué: On the role of knowledge graphs in explainable AI. Semantic Web 11(1): 41-51 (2020)

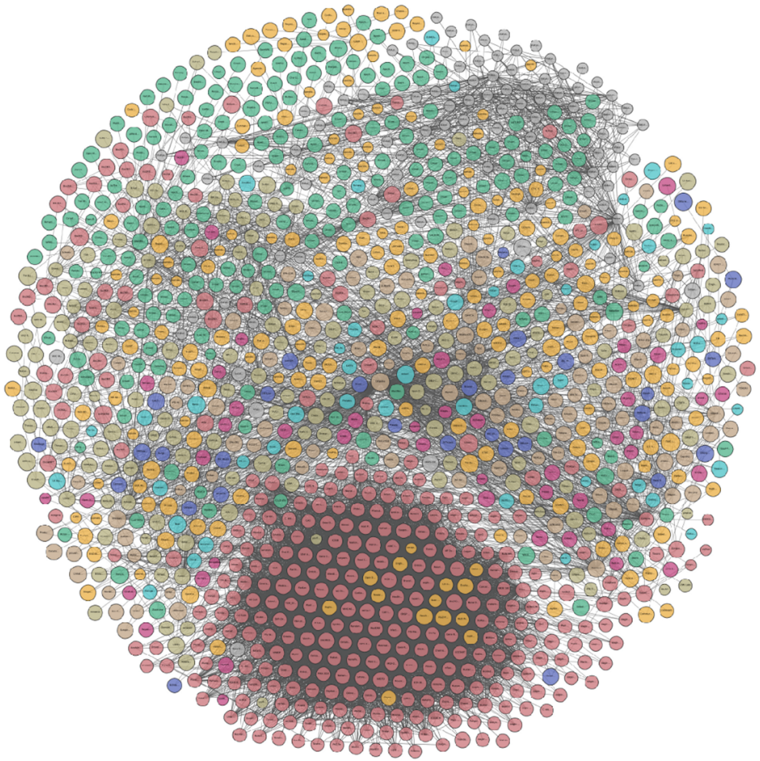
Knowledge Graph (2)

Name	Entities	Relations	Types	Facts
Freebase	40M	35K	26.5K	637M
DBpedia (en)	4.6M	1.4K	735	580M
YAGO3	17M	77	488K	150M
Wikidata	15.6M	1.7K	23.2K	66M
NELL	2M	425	285	433K
Google KG	570M	35K	1.5K	18B
Knowledge Vault	45M	4.5K	1.1K	271M
Yahoo! KG	3.4M	800	250	1.39B

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

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

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



Freddy Lécué: On the role of knowledge graphs in explainable AI. Semantic Web 11(1): 41-51 (2020)

Knowledge Graph Construction

Knowledge Graph construction methods can be classified in:

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

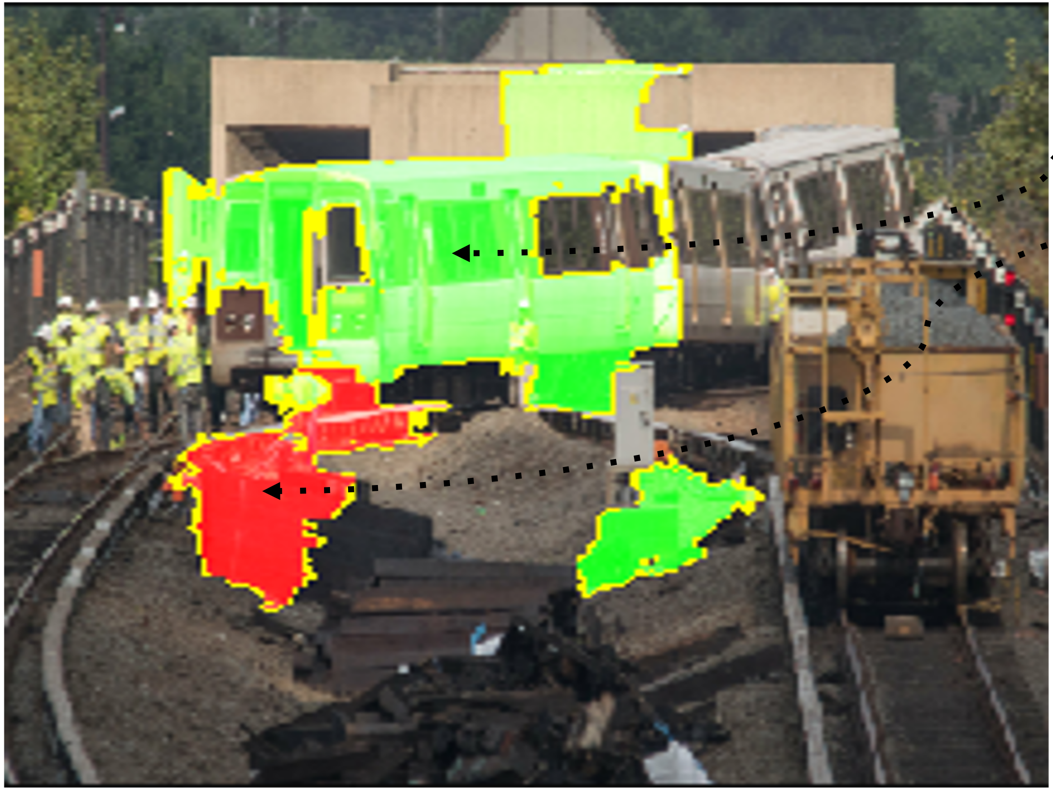
Coverage is an issue:

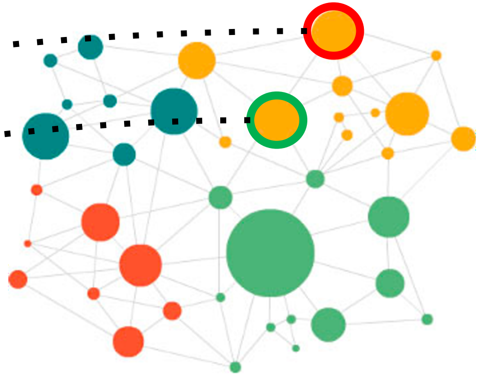
- **Freebase** (40M entities) - 71% of persons without a birthplace, 75% without a nationality, even worse for other relation types [Dong et al. 2014]
- **DBpedia** (20M entities) - 61% of persons without a birthplace, 58% of scientists missing why they are popular [Krompaß et al. 2015]

Relational Learning can help us overcoming these issues.

Freddy Lécué: On the role of knowledge graphs in explainable AI. Semantic Web 11(1): 41-51 (2020)

Knowledge Graph in Machine Learning (1)



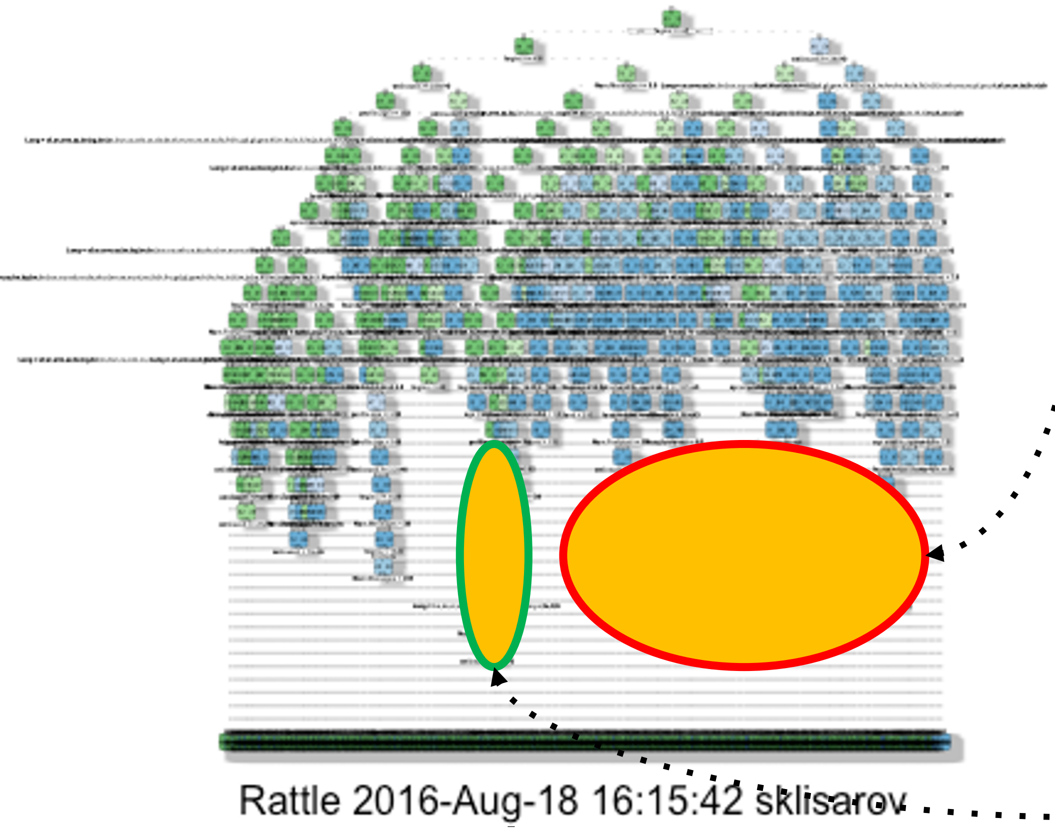


Augmenting (input) features with more semantics such as knowledge graph embeddings / entities

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

Freddy Lécué: On the role of knowledge graphs in explainable AI. Semantic Web 11(1): 41-51 (2020)

Knowledge Graph in Machine Learning (2)

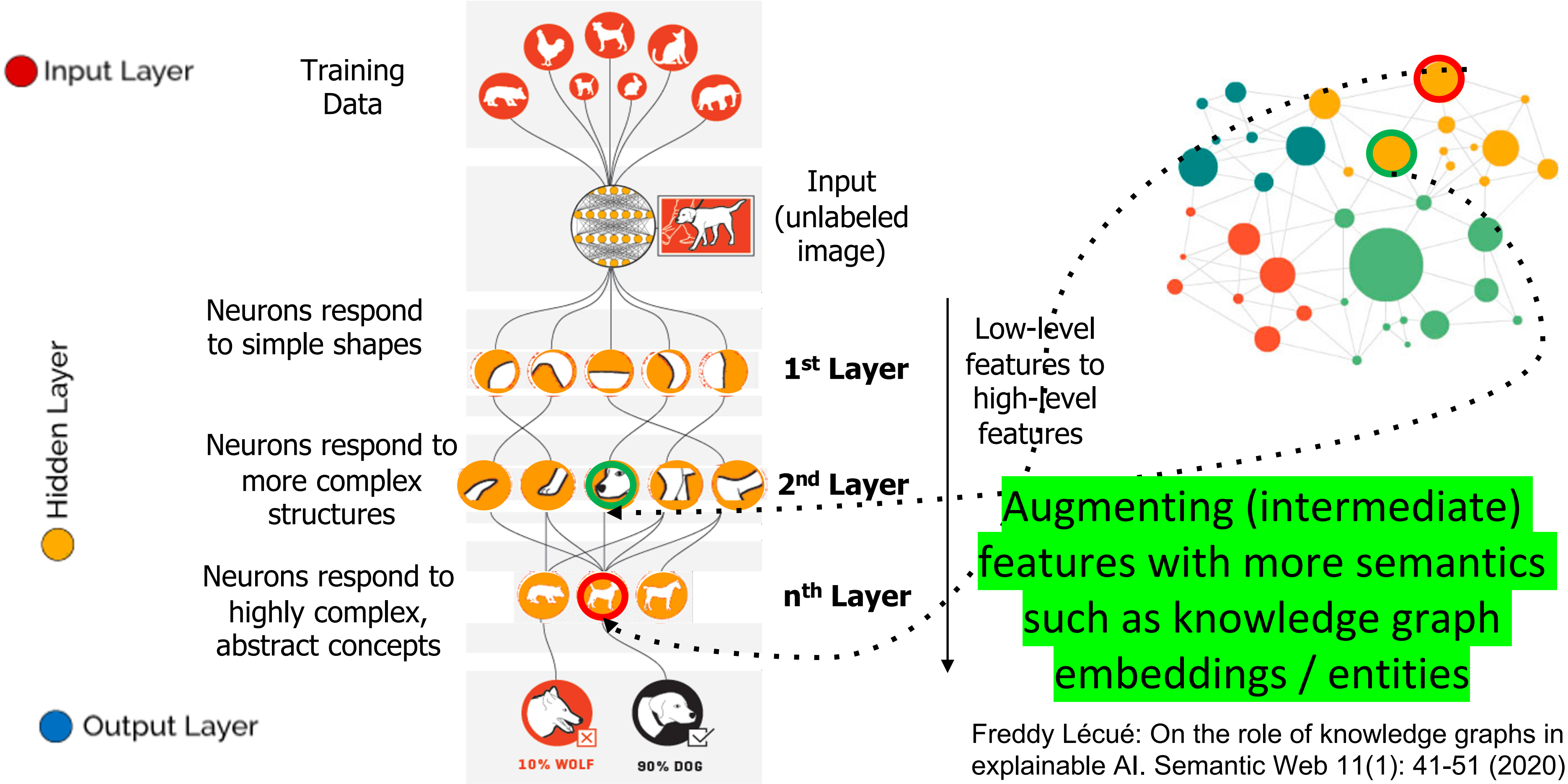


https://stats.stackexchange.com/questions/230581/decision-tree-too-large-to-interpret

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

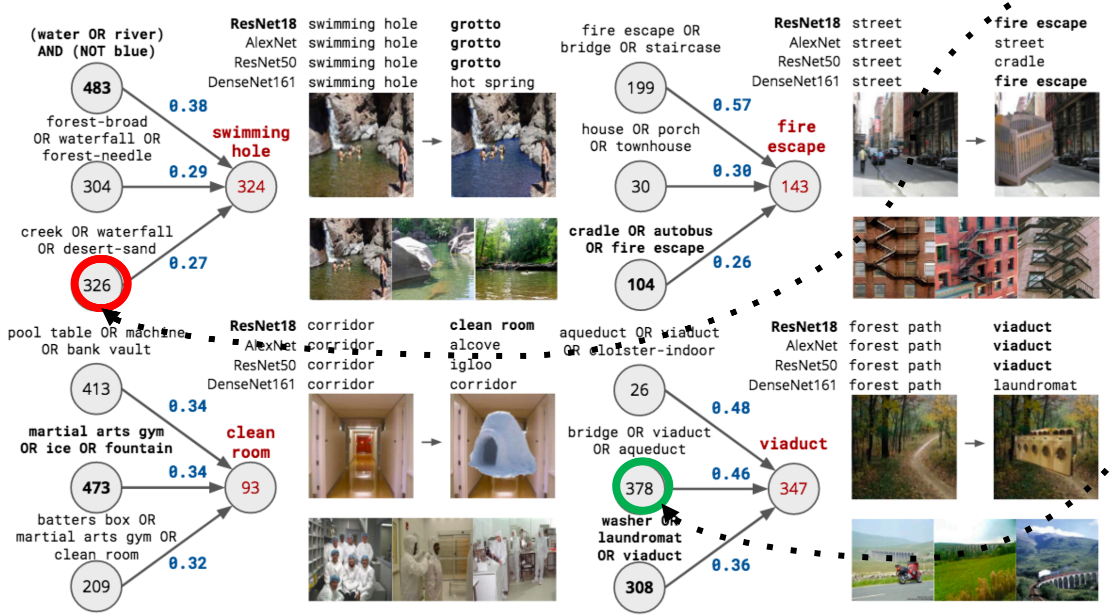
Freddy Lécué: On the role of knowledge graphs in explainable AI. Semantic Web 11(1): 41-51 (2020)

Knowledge Graph in Machine Learning (3)



Freddy Lécué: On the role of knowledge graphs in explainable AI. Semantic Web 11(1): 41-51 (2020)

Knowledge Graph in Machine Learning (4)

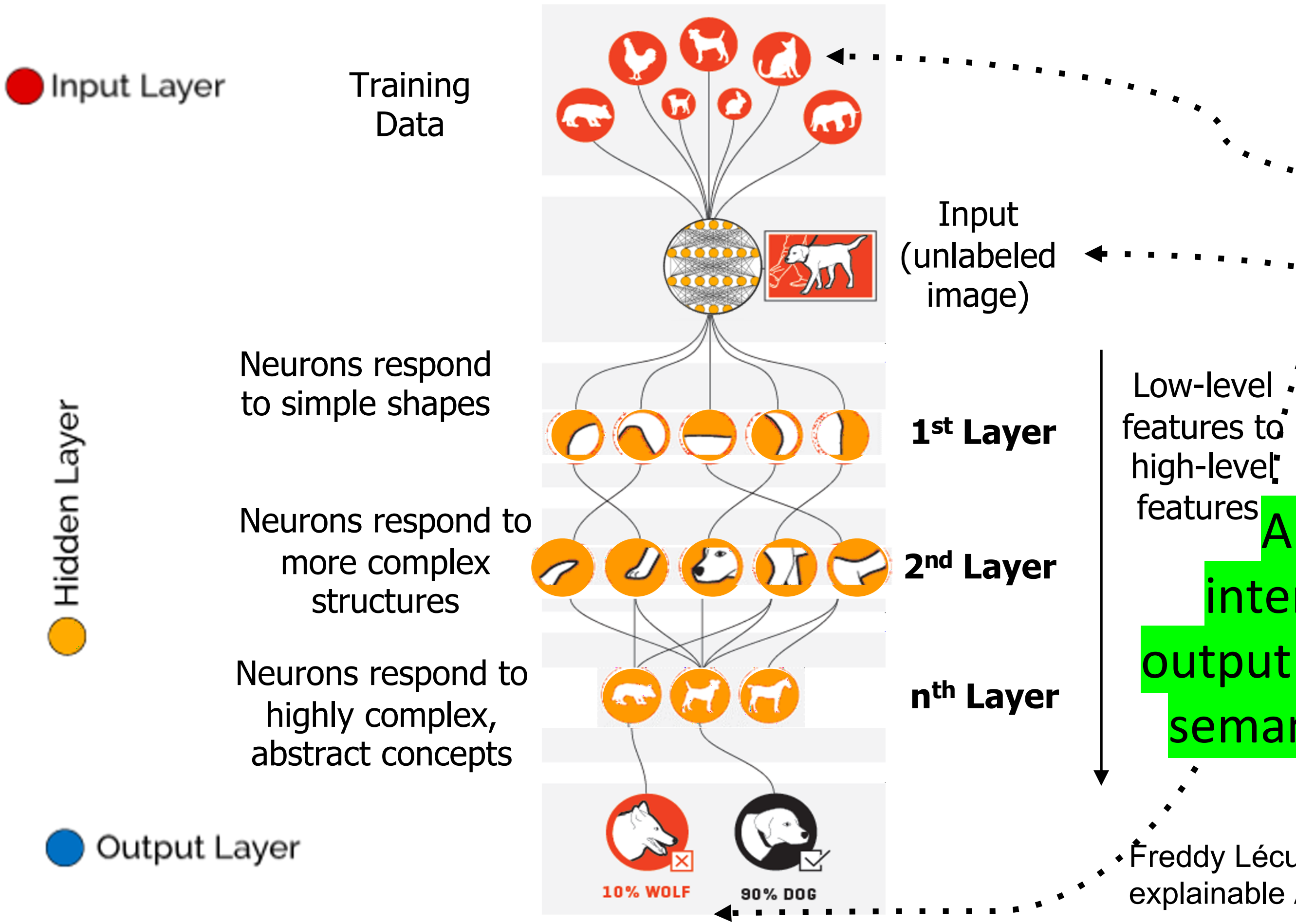


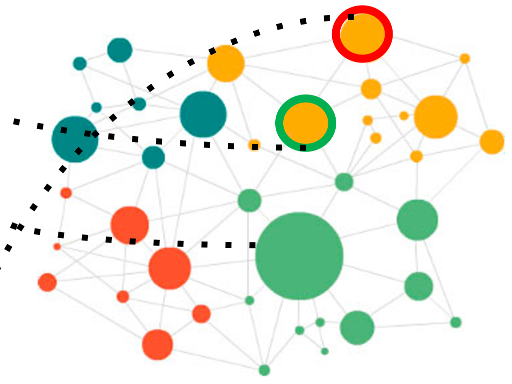
Jesse Mu, Jacob Andreas:Compositional Explanations of Neurons. NeurIPS 2020

Low-level features to high-level features

Open question: What is the impact of semantic representation on units in Neural Networks?

Knowledge Graph in Machine Learning (5)





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

Freddy Lécué: On the role of knowledge graphs in explainable AI. Semantic Web 11(1): 41-51 (2020)

Knowledge Graph in Machine Learning (6)

Description 1: This is an orange train accident

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

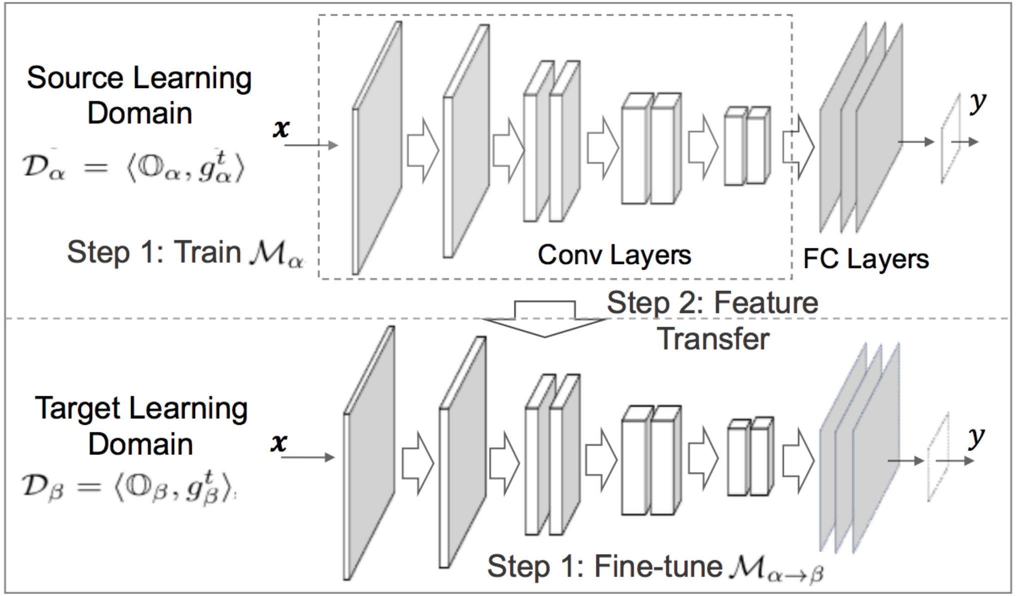
Description 3: This is a public transportation accident

Augmenting models with semantics to support personalized explanation

Freddy Lécué: On the role of knowledge graphs in explainable AI. Semantic Web 11(1): 41-51 (2020)

Knowledge Graph in Machine Learning (7)

"How to explain transfer learning with appropriate knowledge representation?



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

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

Knowledge Graph in Machine Learning (8)

"How to explain concept drift in Machine Learning?

Augmenting input features and domains with semantics to interpret concept drift in Machine Learning

Without semantics augmentation

With semantics augmentation

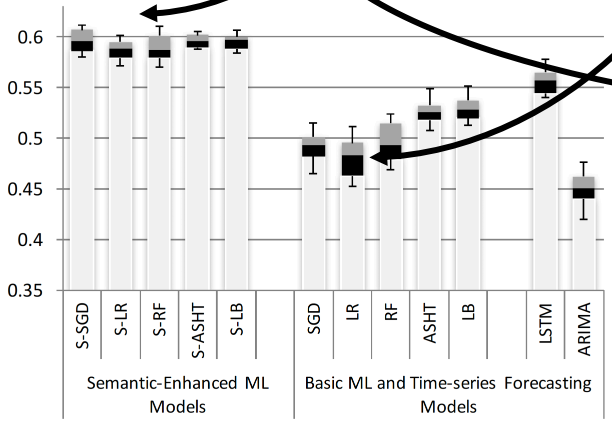


Figure 6: [Beijing Context] Baseline Comparison of Forecasting Macro-F1 Score (Evaluation of Algorithm 1-3), where $\Delta = 6$.

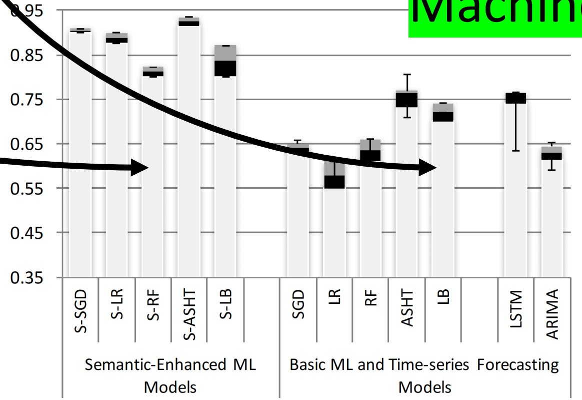


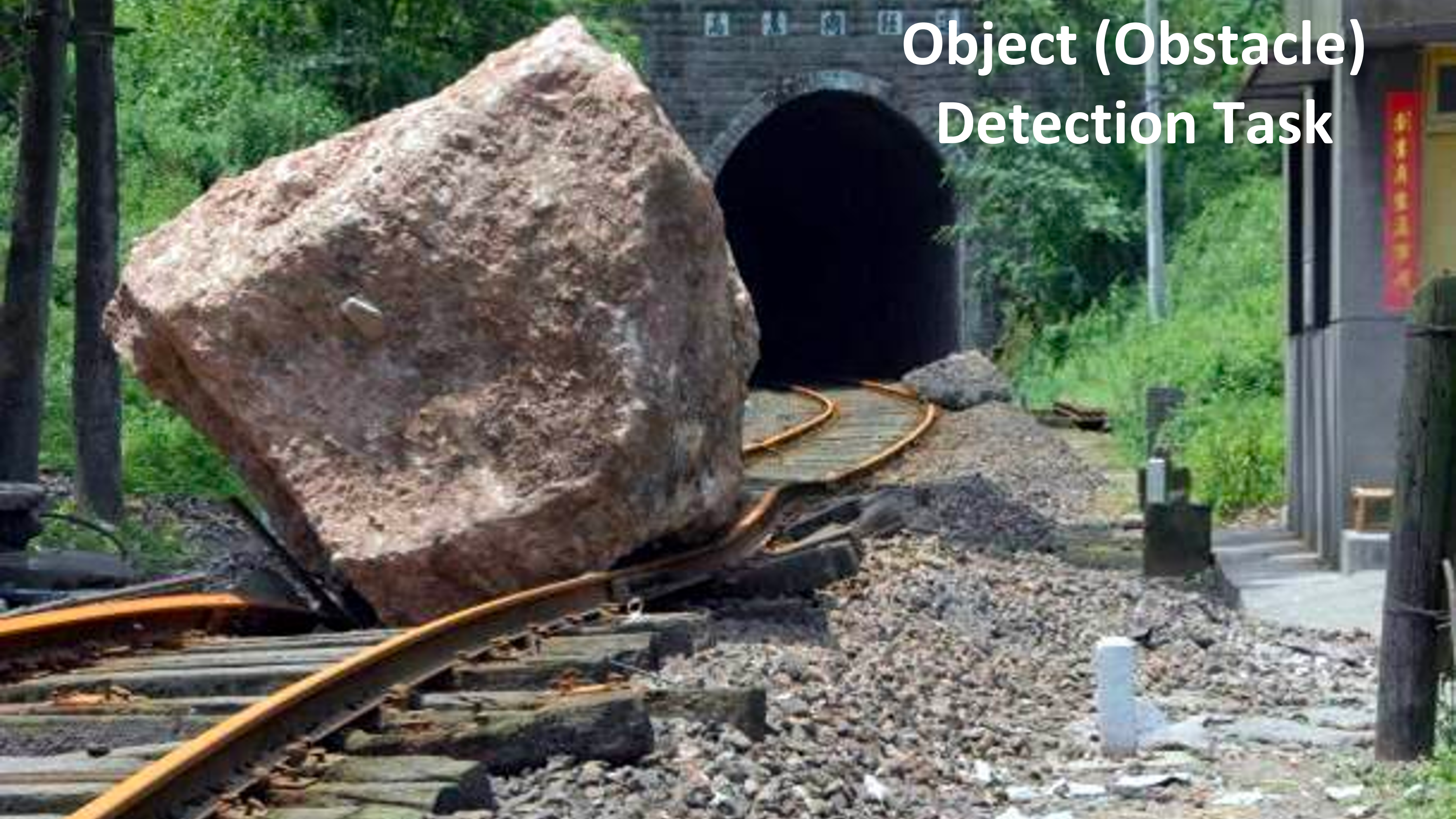
Figure 7: [Dublin Context] Baseline Comparison of Forecasting Macro-F1 Score (Evaluation of Algorithm 1-3), where $\Delta = 6$.

Jiaoyan Chen and Freddy Lécué and Jeff Z. Pan and Shumin Deng and Huajun Chen. Knowledge graph embeddings for dealing with concept drift in machine learning. Journal of Web Semantics. (2021) http://www.sciencedirect.com/science/article/pii/S1570826820300585

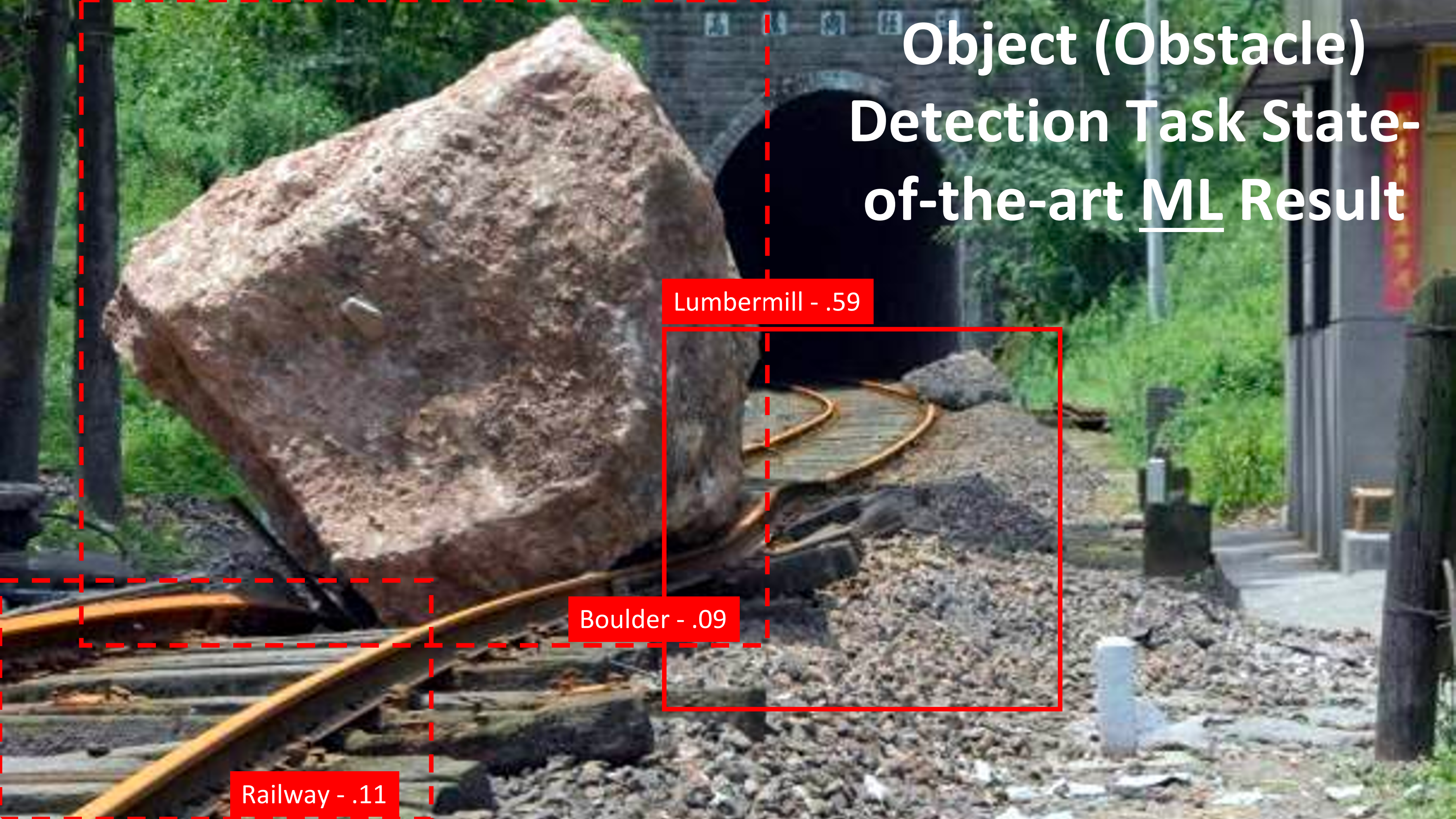
How Does it Work in Practice?

State of the Art Machine Learning Applied to Critical Systems

Object (Obstacle) Detection Task




Object (Obstacle) Detection Task State-of-the-art ML Result
Lumbermill - .59


Object (Obstacle) Detection Task State-of-the-art ML Result
Lumbermill - .59
Boulder - .09
Railway - .11

State of the Art
XAI
Applied to Critical
Systems

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

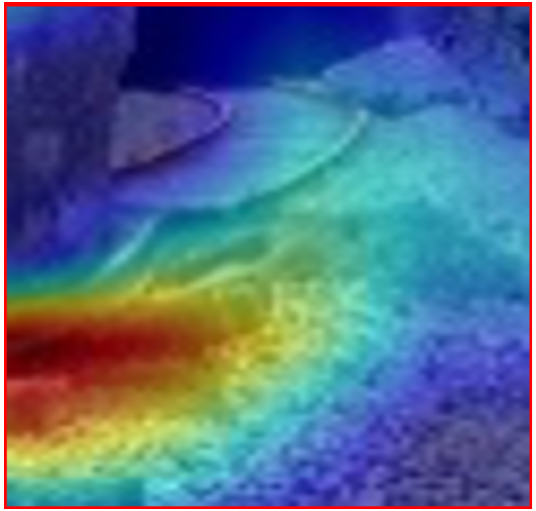
Unfortunately, this is of NO use for a human behind the system

Let’s stay back

Why this Explanation?
(meta explanation)

After Human Reasoning…

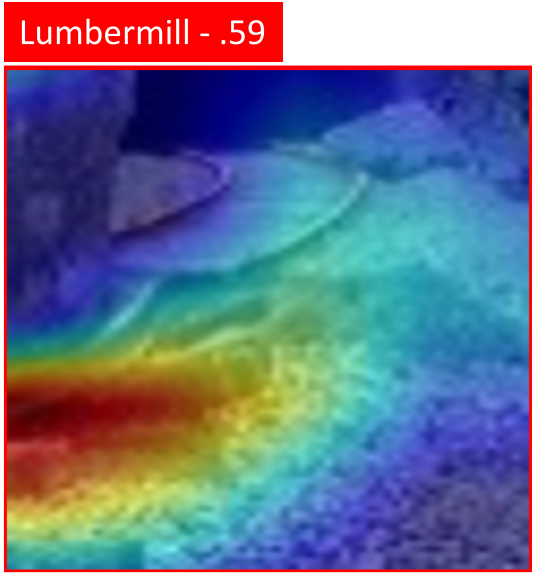
Lumbermill - .59



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

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

What is missing?



DBpedia Browse using Formats Faceted Browser Sparql Endpoint

About: Boulder

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

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

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

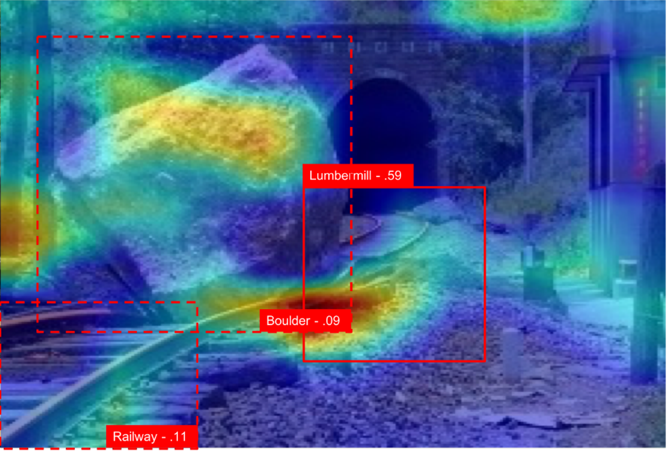
DBpedia Browse using Formats Faceted Browser Sparql Endpoint

About: Rail transport

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

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

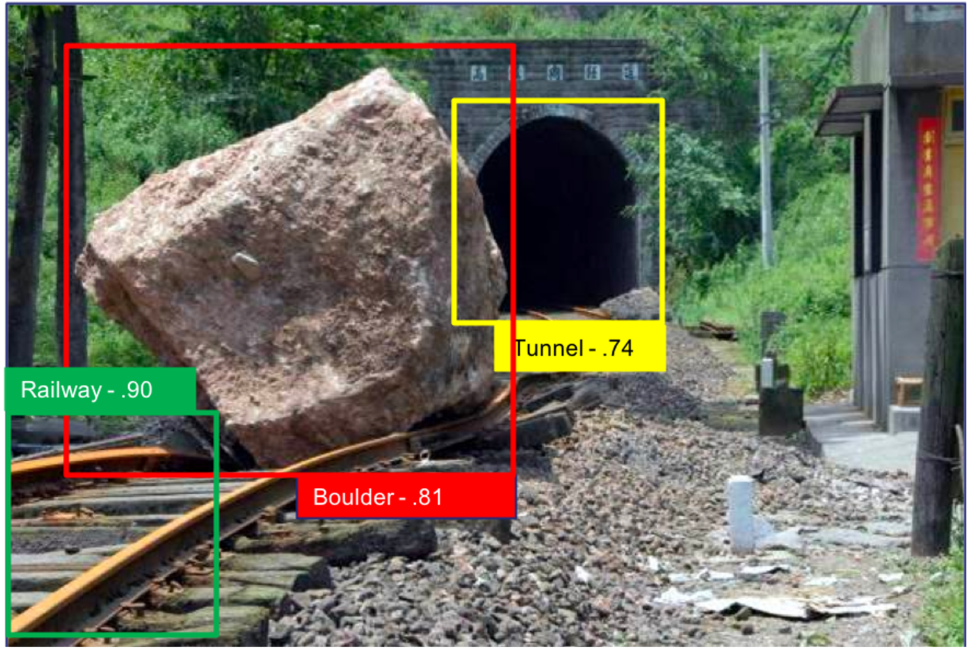
Property	Value
dbo:abstract	• Rail transport is a means of conveyance of passengers and goods on wheeled vehicles running on rails, also known as tracks. It is also commonly referred to as train transport. In contrast to road transport, where vehicles run on a prepared flat surface, rail vehicles (rolling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



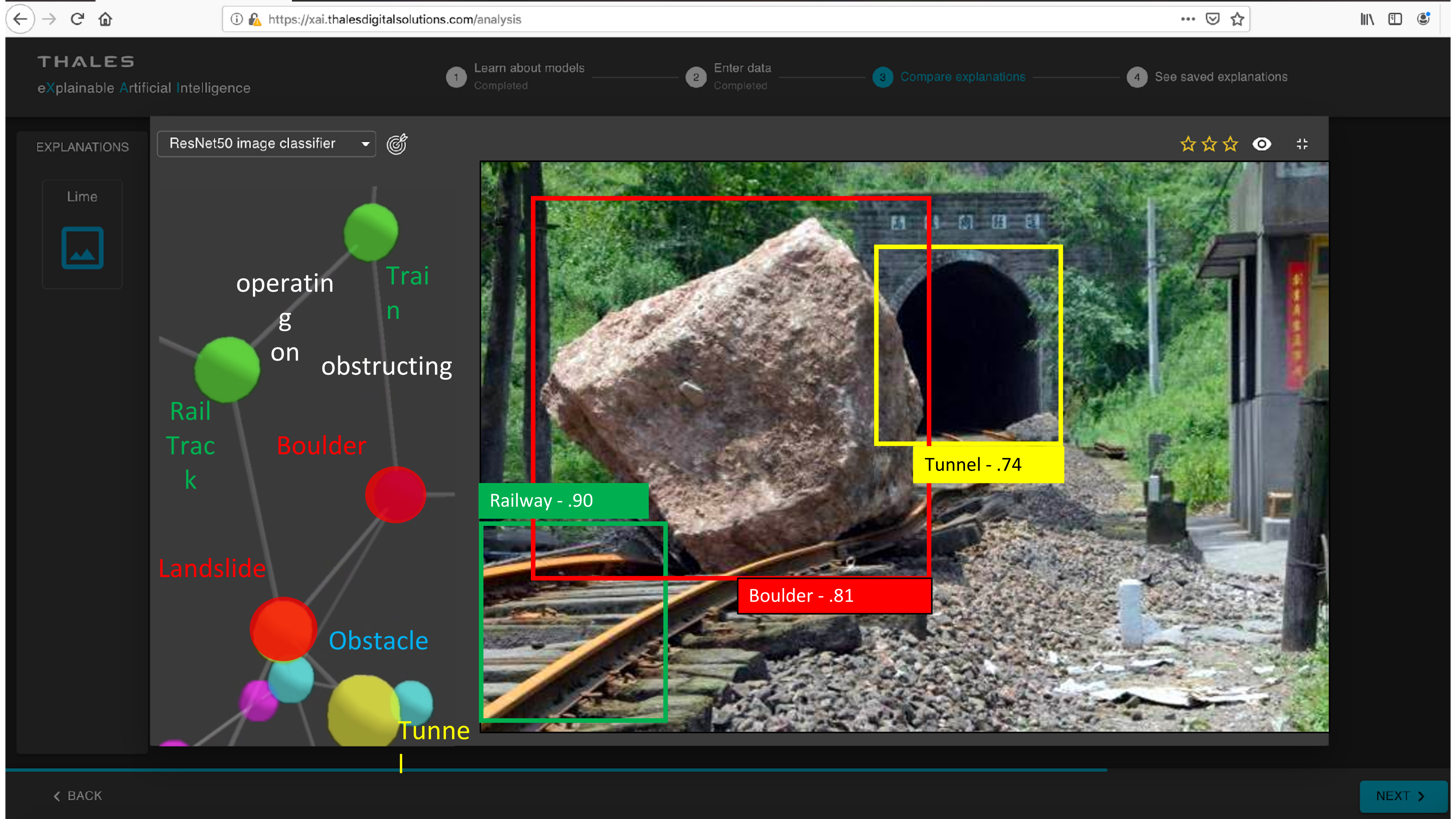
- **Hardware**: **High performance, scalable, generic** (to different FGPA family) **& portable** CNN dedicated **programmable** processor implemented on an FPGA for **real-time embedded inference**

- **Software**: Knowledge graph extension of object detection

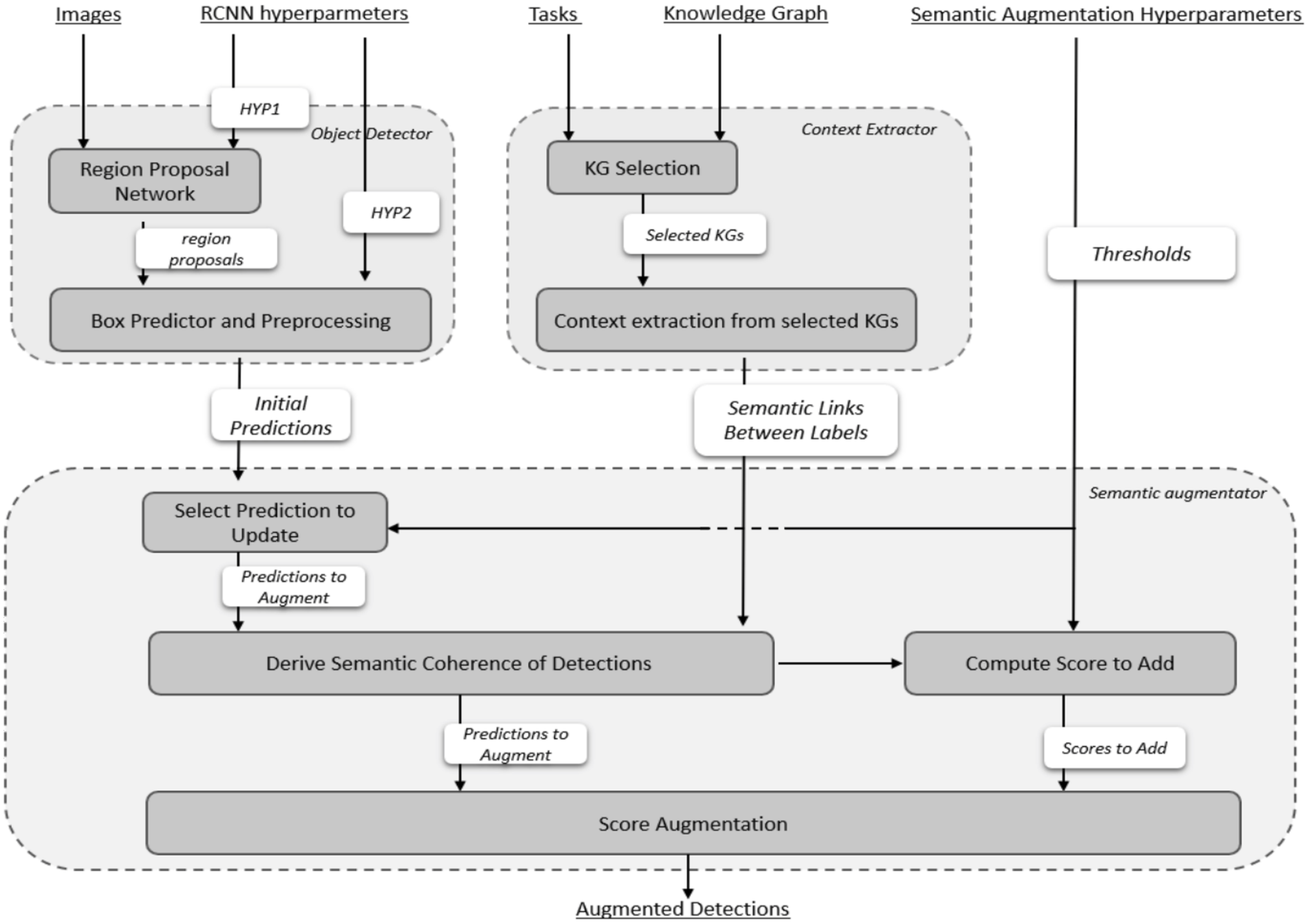
Transitionin g



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


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

Knowledge Graph in Machine Learning - An Implementation



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

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

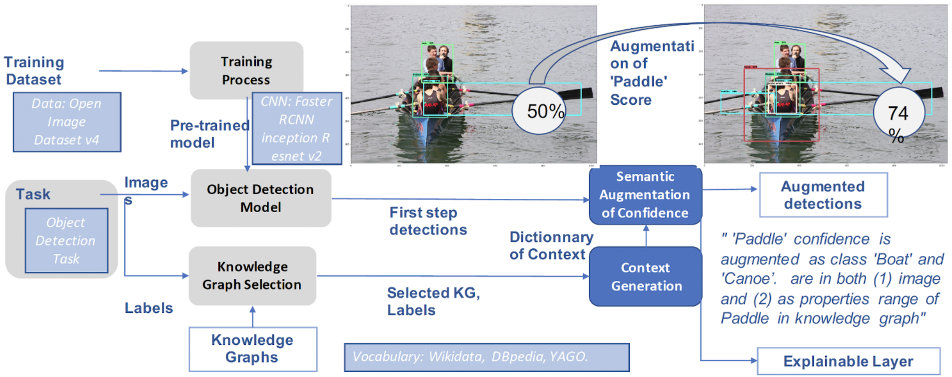
Freddy Lécué, Baptiste Abeloos, Jonathan Anctil, Manuel Bergeron, Damien Dalla-Rosa, Simon Corbeil-Letourneau, Florian Martet, Tanguy Pommellet, Laura Salvan, Simon Veilleux, Maryam Ziaeefard: Thales XAI Platform: Adaptable Explanation of Machine Learning Systems - A Knowledge Graphs Perspective. ISWC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

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

Part IV

XAI Applications and Lessons Learnt

Explainable Boosted Object Detection – Industry Agnostic



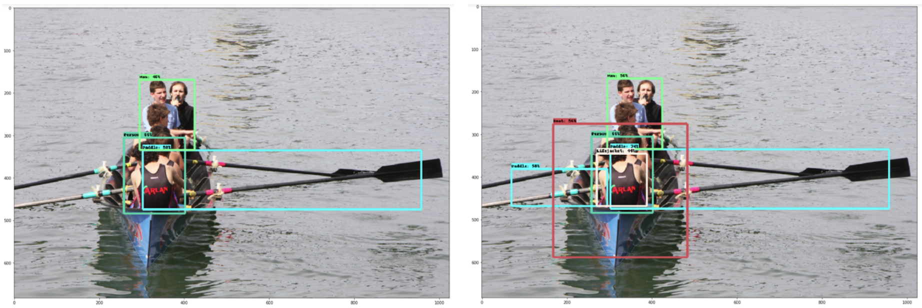


Fig. 2. Left image: results from baseline Faster RCNN: Paddle: 50% confidence, Person: 66%, Man: 46%. Right image: results from the semantic augmentation: **Paddle: 74%** confidence, Person: 66%, **Man: 56%**, **Boat: 58% with explanation: Person, Paddle, Water as part of the context in the image and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

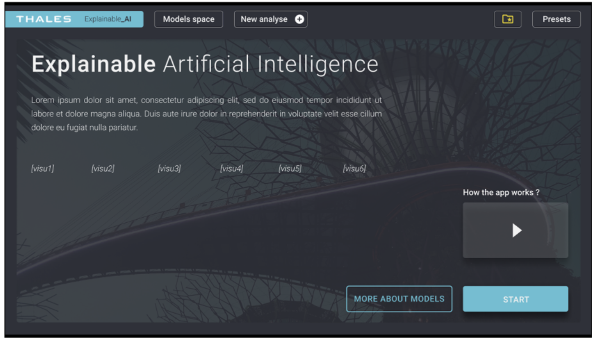
Challenge: Object detection is usually performed from a large portfolio of Artificial Neural Networks (ANNs) architectures trained on large amount of labelled data. Explaining object detections is rather difficult due to the high complexity of the most accurate ANNs.

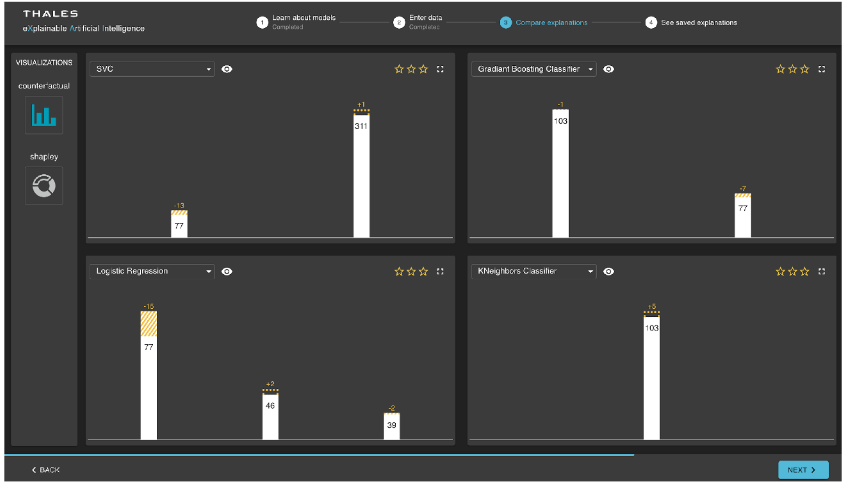
AI Technology: Integration of AI related technologies i.e., Machine Learning (Deep Learning / CNNs), and knowledge graphs / linked open data.

XAI Technology: Knowledge graphs and Artificial Neural Networks

THALES

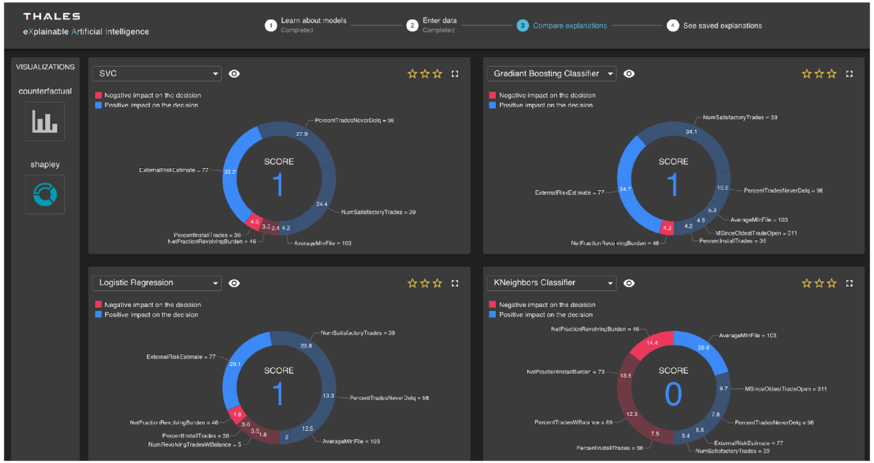
Thales XAI Platform – Industry Agnostic





Context

- Explanation in Machine Learning systems has been identified to be the one asset to have for large scale deployment of Artificial Intelligence (AI) in critical systems
- Explanations could be example-based (who is similar), features-based (what is driving decision), or even counterfactual (what-if scenario) to potentially action on an AI system; they could be represented in many different ways e.g., textual, graphical, visual



Goal

- All representations serve different means, purpose and operators. We designed the first-of-its-kind XAI platform for critical systems i.e., the Thales Explainable AI Platform which aims at serving explanations through various forms

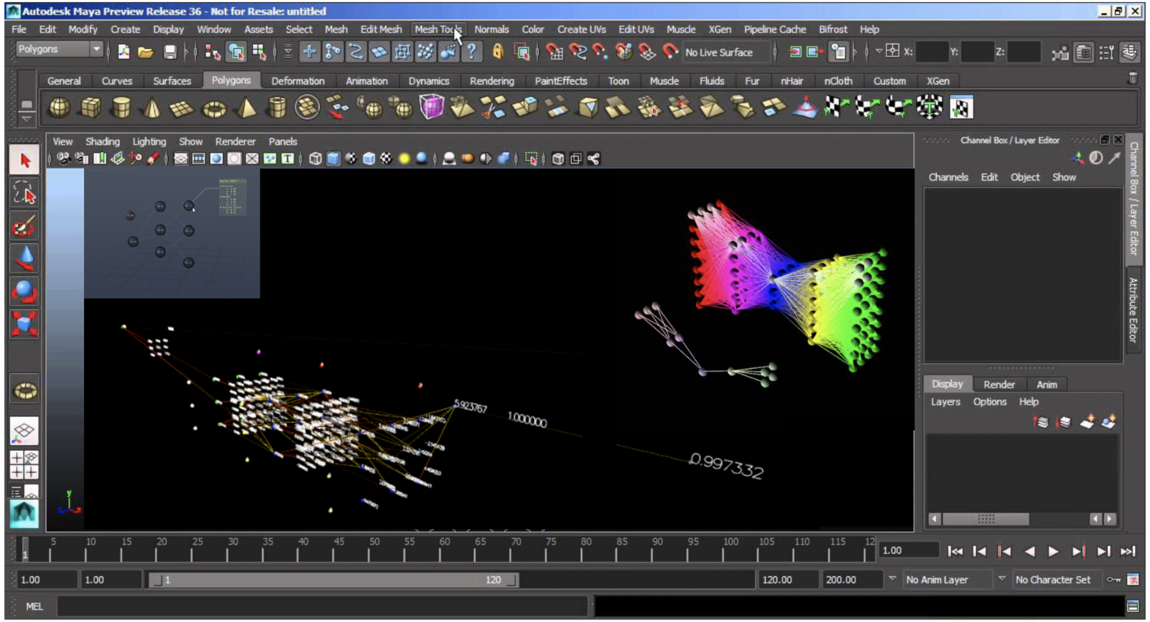
Approach: Model-Agnostic

- [AI:ML] Grad-Cam, Shapley, Counter-factual, Knowledge graph

THALES

Video: https://drive.google.com/file/d/1zoKidieGH5zaahOn8ekXXBo74BEeZvc-/view

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

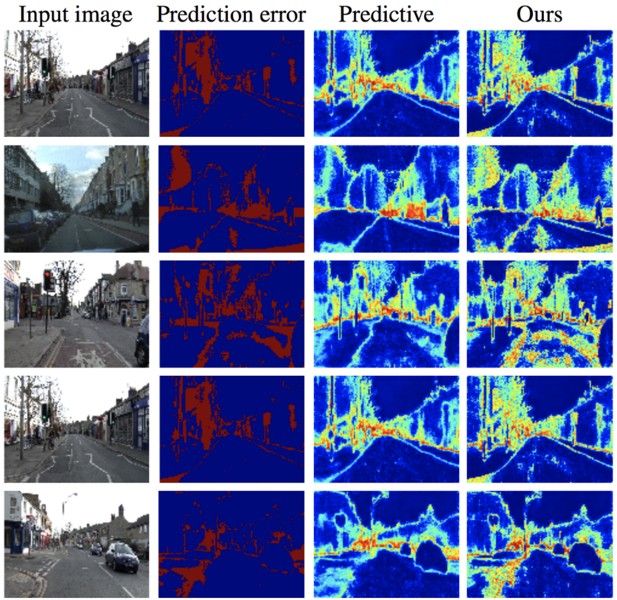
XAI Technology: Artificial Neural Network, 3D Modeling and Simulation Platform For AI

Video: https://drive.google.com/file/d/1ZTwndNzC9bN9ouP9cjjuXcyzZ3OYIcgU/view

Zetane.com

Obstacle Identification Certification (Trust) – Transportation

THALES



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

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

XAI Technology: Deep learning and Epistemic uncertainty

Explaining Flight Performance – Transportation

Challenge: Predicting and explaining aircraft engine performance

AI Technology: Artificial Neural Networks

XAI Technology: Shapely Values

Explainable On-Time Performance – Transportation

KLM / Transavia Flight Delay Prediction

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

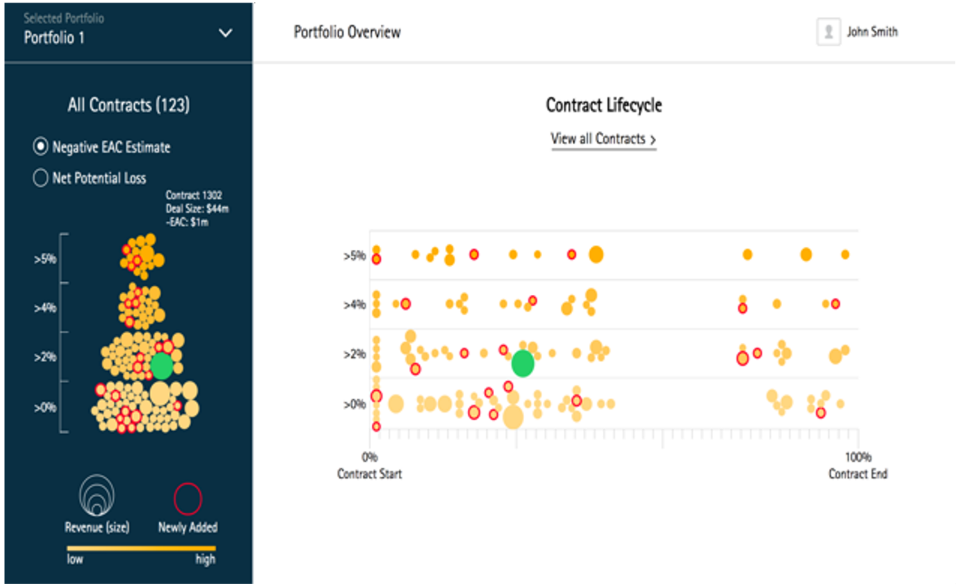
AI Technology: Integration of AI related technologies i.e., Machine Learning (Deep Learning / Recurrent neural Network), Reasoning (through semantics-augmented case-based reasoning) and Natural Language Processing for building a robust model which can (1) predict flight delays in minutes, (2) explain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

XAI Technology: Knowledge graph embedded Sequence Learning using LSTMs

INNOVATION ARCHITECTURE: ACCENTURE LABS

THALES

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

Alvaro H. C. Correia, Freddy Lécué: Human-in-the-Loop Feature Selection. AAAI 2019: 2438-2445

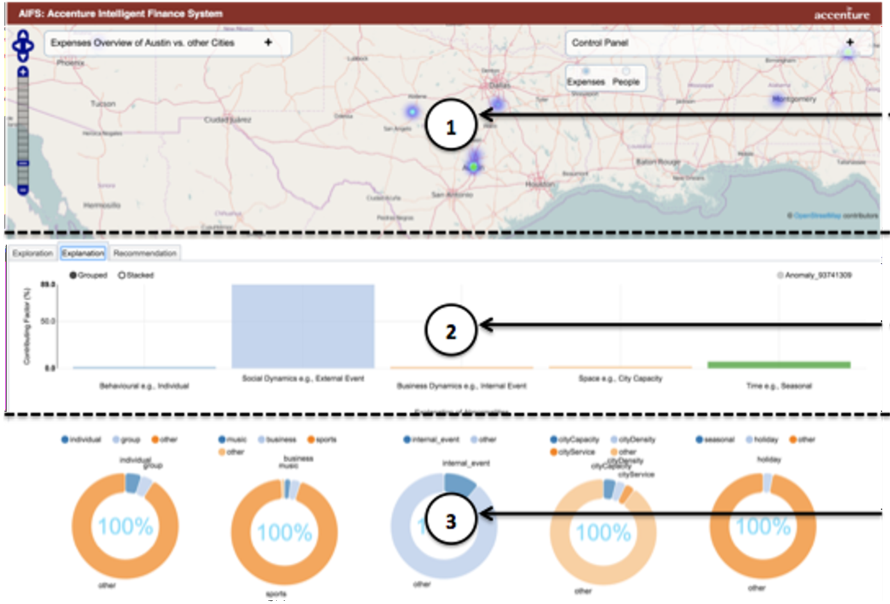
INNOVATION ARCHITECTURE: ACCENTURE LABS

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

Explainable Anomaly Detection – Finance (Compliance)

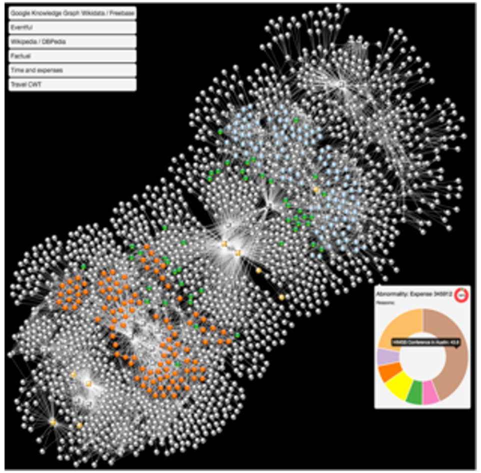


INNOVATION ARCHITECTURE:
ACCENTURE LABS

Data analysis for spatial interpretation of abnormalities: abnormal expenses

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

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



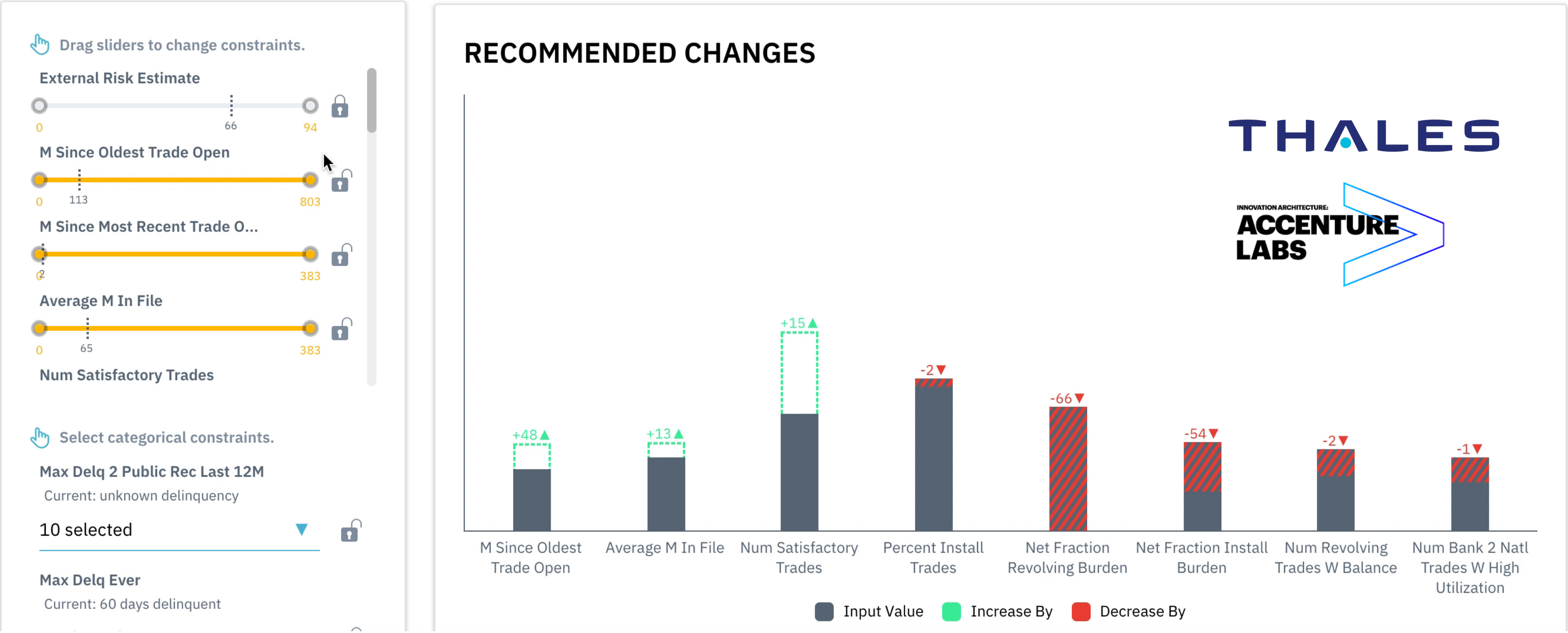
Freddy Lécué, Jiewen Wu: Explaining and predicting abnormal expenses at large scale using knowledge graph based reasoning. J. Web Sem. 44: 89-103 (2017)

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

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

XAI Technology: Knowledge graph embedded Ensemble Learning . **Video**: https://www.dropbox.com/s/sst232gu0yeqy21/IUI-2017-Final.mp4?dl=0

Counterfactual Explanations for Credit Decisions – 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.

Explanation of Medical Condition Relapse – Health

THALES

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

Explaining Visual Question Answering – Industry Agnostic

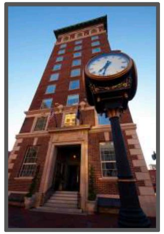
Tabular QA

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

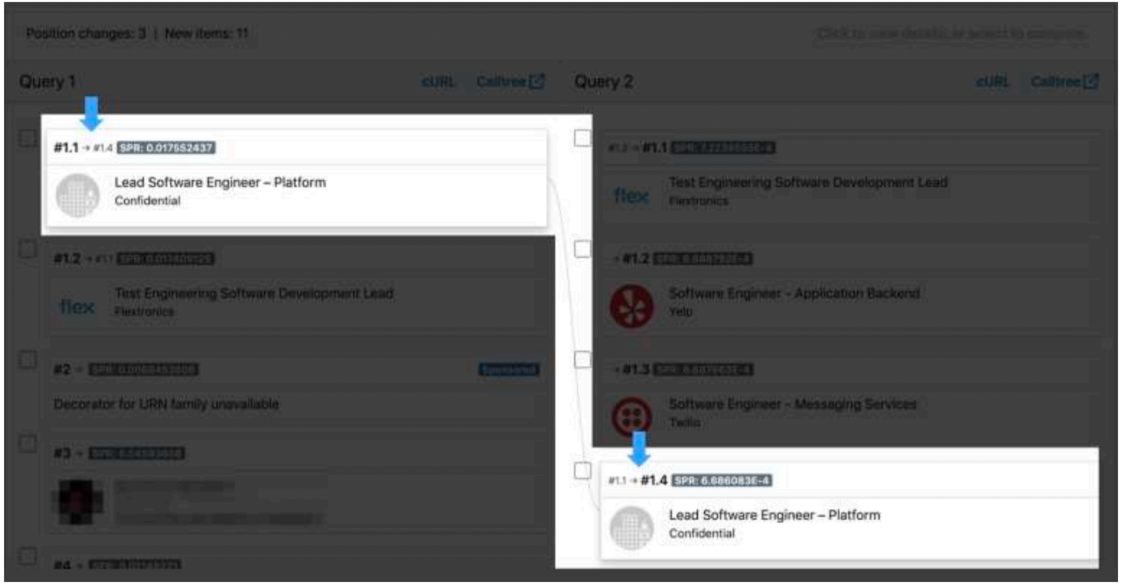
Source: Explainable AI in Industry. KDD 2019 Tutorial. Ankur Taly, Mukund Sundararajan, Kedar Dhamdhere, Pramod Mudrakarta

Relevance Debugging and Explaining – Industry Agnostic

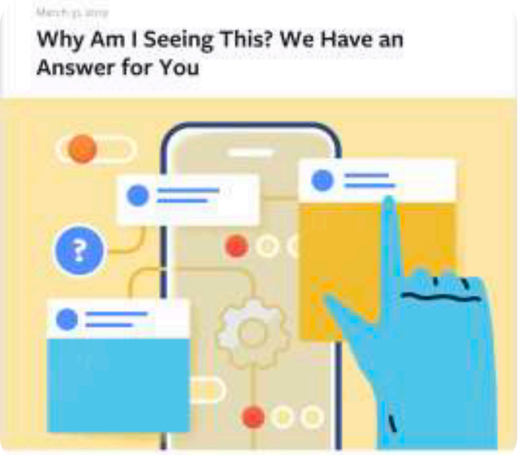
Challenge: A Machine Learning system can fail in many different points e.g., data features selection, construction, inconsistencies. How to debug bad performance in machine learning models and prediction?

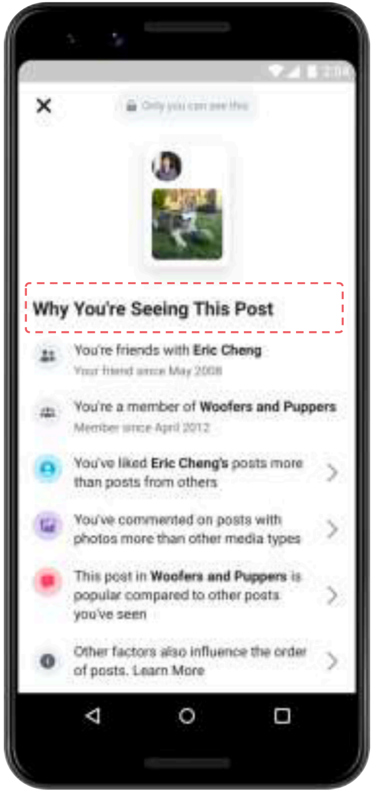
AI Technology: Artificial Neural Networks.

XAI Technology: Model / Prediction comparison



Explaining Recommendation – Social Media



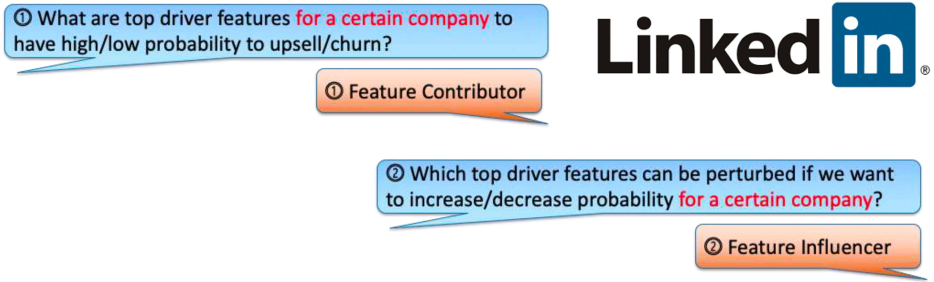


Challenge: How to establish trust between Social Media and their users? Explaining post / news recommendation is crucial for users to engage with content providers.

AI Technology: Artificial Neural Networks.

XAI Technology: Recommendation-based

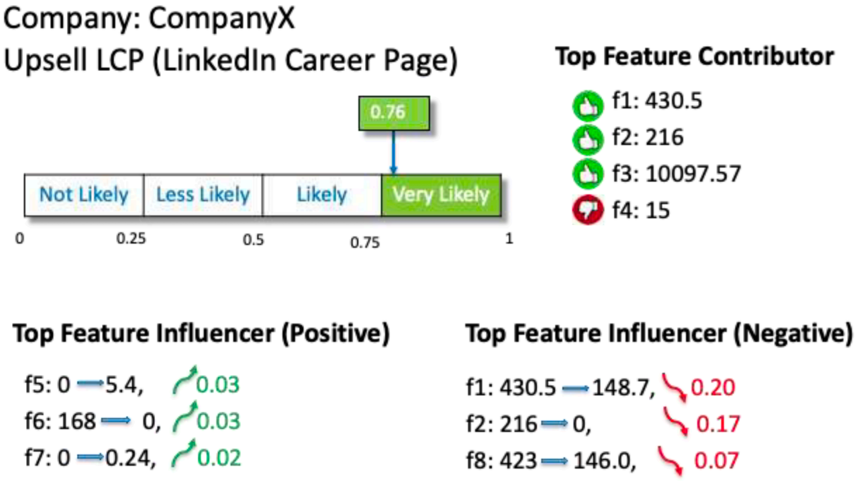
Model Explanation for Sales Prediction – Sales



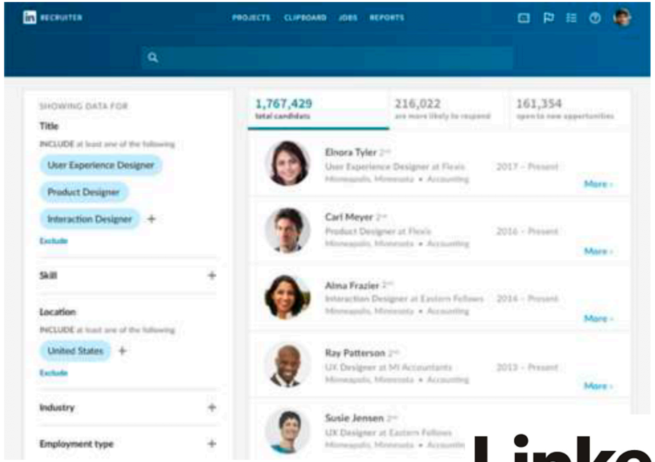
Challenge: How to predict and explain upsell / churn for a company?

AI Technology: Artificial Neural Networks.

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



Explaining Talent Search Results – Human Resources

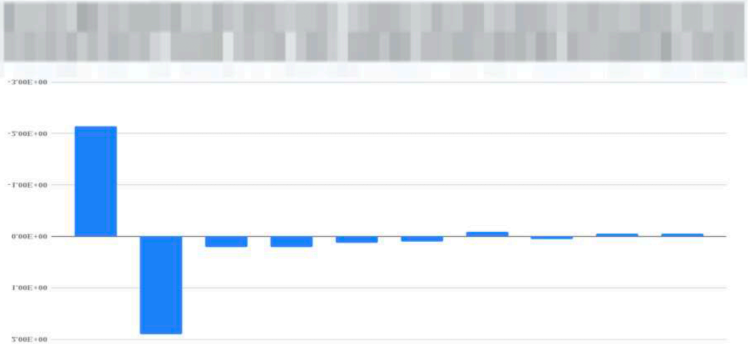


Challenge: How to rationalize a talent search for a recruiter when looking for candidates for a given role. Features are dynamic and costly to compute. Recruiters are interested in discriminating between two candidates to make a selection.

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

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

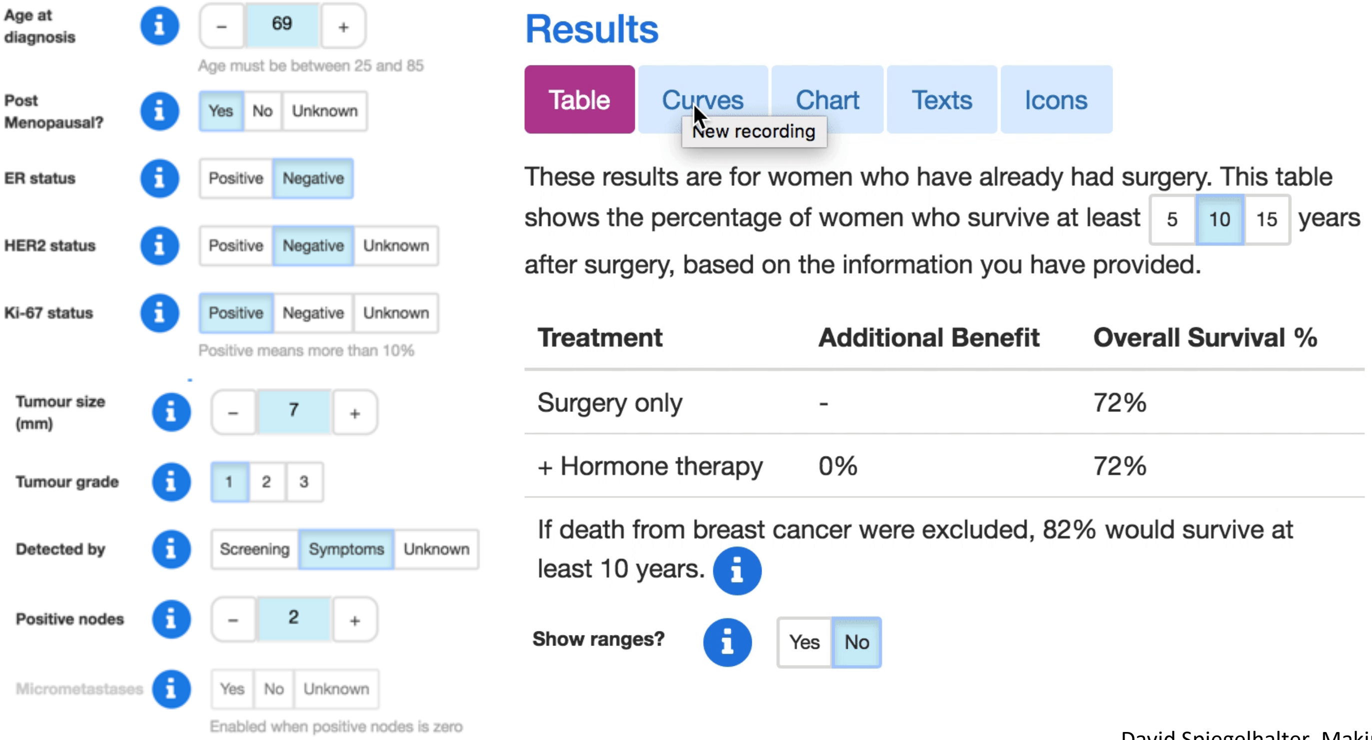
LinkedIn



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

Explaining Breast Cancer Survival Rate Prediction – Health

predict
breast cancer



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.

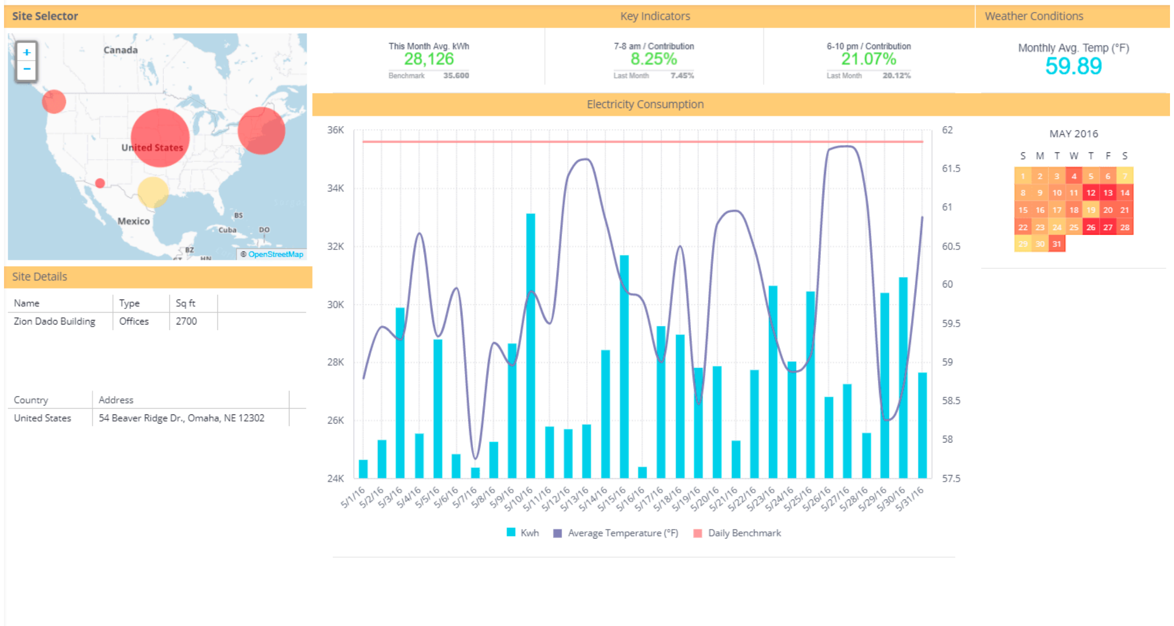
AI Technology: competing risk analysis

XAI Technology: Interactive explanations, Multiple representations.

David Spiegelhalter, Making Algorithms trustworthy, NeurIPS 2018 Keynote

predict.nhs.uk/tool

Explaining Energy Consumption – A Global Perspective – Energy

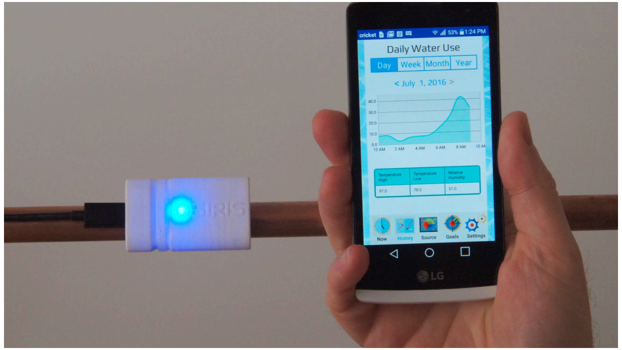


Challenge: Predicting energy consumption is crucial to satisfy high-demand. However some demands might be difficult to forecast, particularly in case of abnormal events. How to augment energy consumption data with open / event data to reach better accuracy and explainability of out-of-distribution demand.

AI Technology: Artificial Neural Network

XAI Technology: Artificial Neural Network, Data Augmentation, Knowledge Graphs

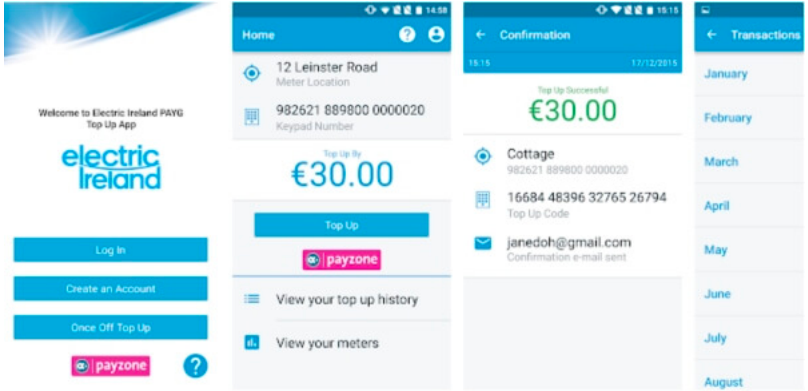
Explaining Energy Consumption – A Local Perspective – Energy



Challenge: Predicting local (home) energy consumption is crucial to satisfy high-demand. Local understanding of consumption requires high-granularity data about energy consumption, which is achieved by analyzing energy signature, and characterizing user patterns on energy consumption.

AI Technology: Artificial Neural Network

XAI Technology: Artificial Neural Network, Shapley values



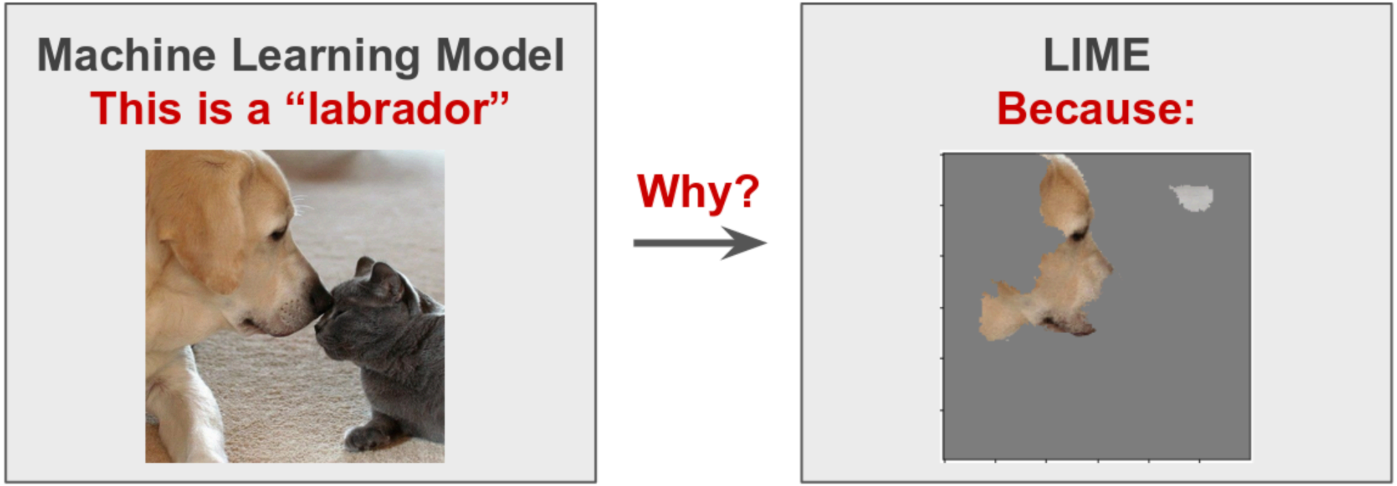
electric Ireland

Part V

XAI Tools, Coding Practices,

Conclusion, and Research Challenges

XAI LIME on Image – Local Input Exploration



In this post, we will study how LIME (Local Interpretable Model-agnostic Explanations) (Ribeiro et. al. 2016) generates explanations for image classification tasks. The basic idea is to understand why a machine learning model (deep neural network) predicts that an instance (image) belongs to a certain class (labrador in this case). For an introductory guide about how LIME works, I recommend you to check my previous blog post Interpretable Machine Learning with LIME. Also, the following YouTube video explains this notebook step by step.

http://t.ly/c3yz

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

XAI LUCID on Image – Neurons Exploration

Lucid: A Quick Tutorial

This tutorial quickly introduces **Lucid**, a network for visualizing neural networks. Lucid is a kind of spiritual successor to DeepDream, but provides flexible abstractions so that it can be used for a wide range of interpretability research.

Note: The easiest way to use this tutorial is as a colab notebook, which allows you to dive in with no setup. We recommend you enable a free GPU by going:

> **Runtime → Change runtime type → Hardware Accelerator: GPU**

Thanks for trying Lucid!

http://t.ly/QqxZ

https://github.com/tensorflow/lucid/
https://distill.pub/2020/circuits/zoom-in/
https://microscope.openai.com/models

XAI GAN Dissection on Image – Network Dissection

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

http://t.ly/x4IF

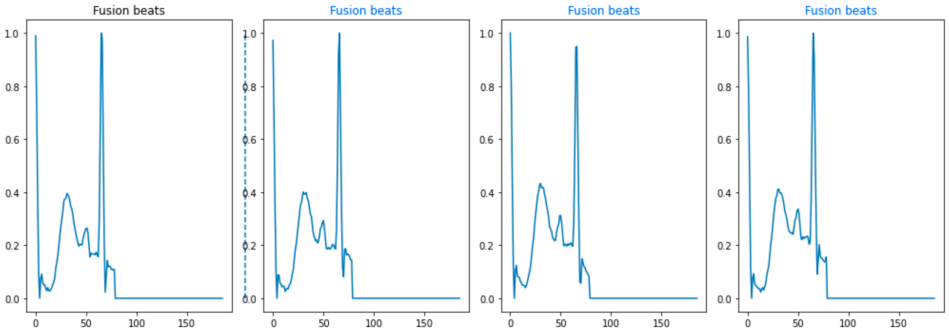
XAI Example-based on Image | Text | EGC – ExMatchina (NeurIPS 2020)

Text
http://t.ly/PNE3

```
negative
18431 REVIEW: you keep disappearing and it makes me a sad panda
18431 Example 1: the end of him and me. very sad ending.
18431 Example 2: Of to work, going to be a very sad day
18431 Example 3: yeah so its been half an hour and still no reply
```

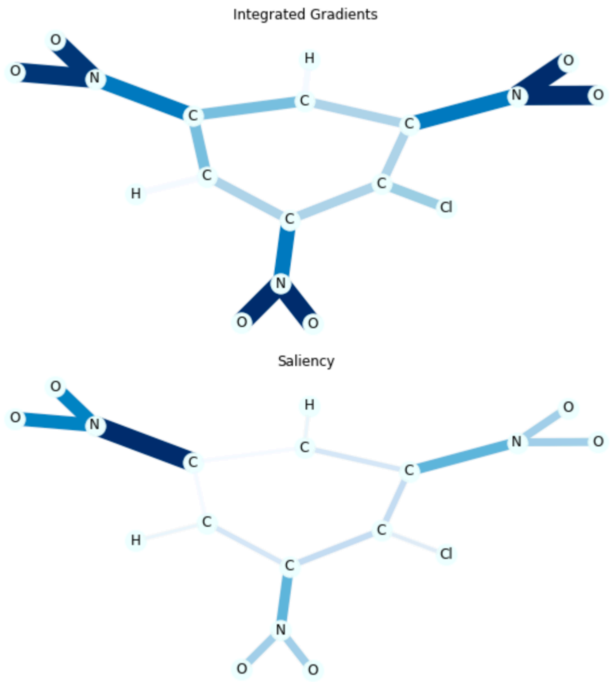
Image
http://t.ly/Jw6L

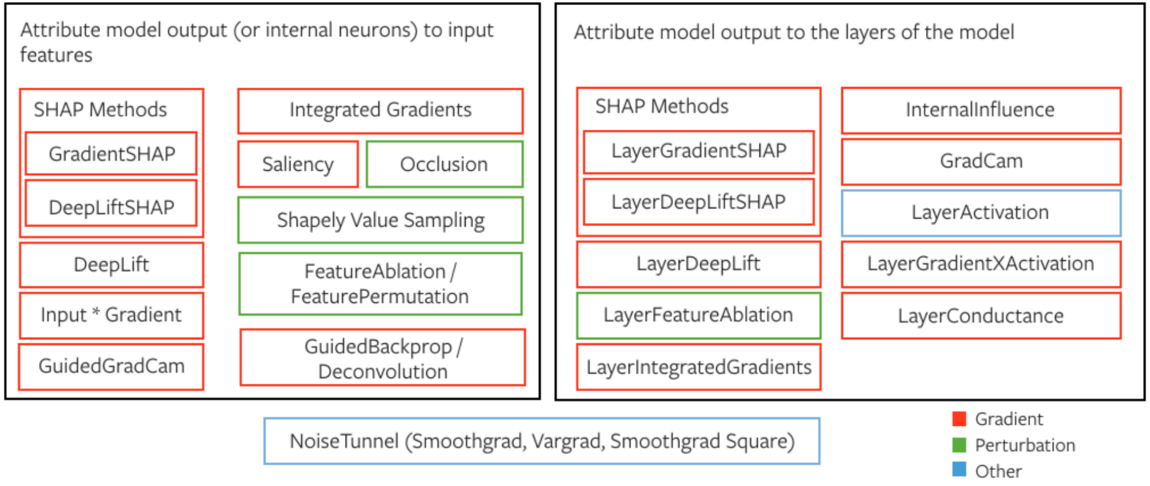
ECG
http://t.ly/EvYG



Jeya Vikranth Jeyakumar, Joseph Noor, Yu-Hsi Cheng, Luis Garcia, Mani B. Srivastava:How Can I Explain This to You? An Empirical Study of Deep Neural Network Explanation Methods. NeurIPS 2020

XAI Integrated Gradient on Graph - Facebook Captum





https://medium.com/pytorch/introduction-to-captum-a-model-interpretability-library-for-pytorch-d236592d8afa

https://captum.ai/

Narine Kokhlikyan, Vivek Miglani, Miguel Martin, Edward Wang, Bilal Alsallakh, Jonathan Reynolds, Alexander Melnikov, Natalia Kliushkina, Carlos Araya, Siqi Yan, Orion Reblitz-Richardson:Captum: A unified and generic model interpretability library for PyTorch. CoRR abs/2009.07896 (2020)

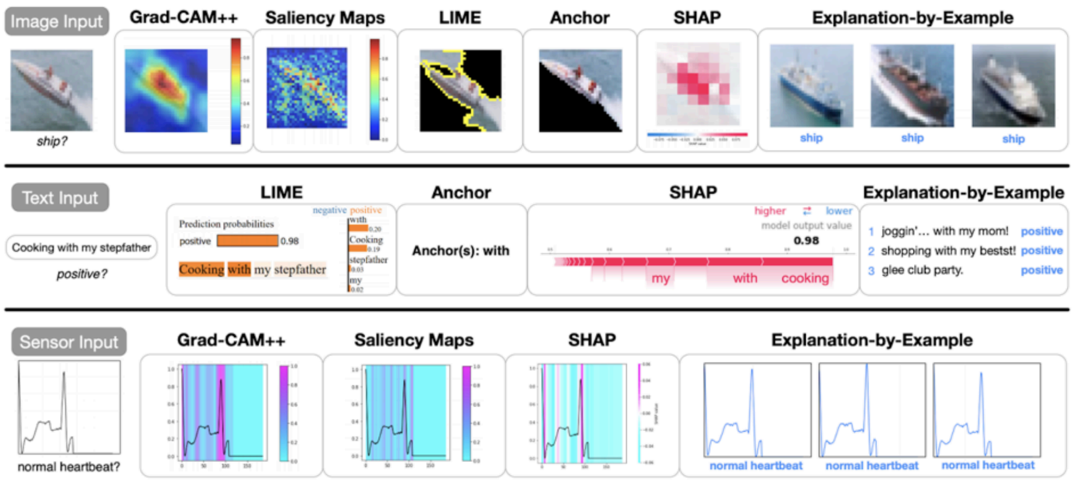
http://t.ly/qMzm

Explanation Comparison

http://t.ly/5nab

Jeya Vikranth Jeyakumar, Joseph Noor, Yu-Hsi Cheng, Luis Garcia, Mani B. Srivastava: How Can I Explain This to You? An Empirical Study of Deep Neural Network Explanation Methods. NeurIPS 2020

https://github.com/nesl/Explainability-Study



Explanation Method	Image Study	Text Study	Audio Study	ECG Study
LIME	47.7 ± 4.5%	**70.4 ± 3.6%**	-	-
Anchor	38.9 ± 4.3%	25.8 ± 3.5%	-	-
SHAP	33.7 ± 4.3%	59.9 ± 3.8%	34.7 ± 4.8%	32.8 ± 3.3%
Saliency Maps	39.4 ± 4.3%	-	46.1 ± 5.1%	40.4 ± 3.5%
GradCAM++	50.8 ± 4.5%	-	48.1 ± 5.3%	42.0 ± 3.5%
Explanation by Examples	**89.6 ± 2.6%**	43.7 ± 3.9%	**70.9 ± 4.7%**	**84.8 ± 2.5%**

More on XAI

Some Tutorials, Workshops, Challenges

Tutorial:

- AAAI 2021 Explainable AI for Societal Event Predictions: Foundations, Methods, and Applications (#1) https://yue-ning.github.io/aaai-21-tutorial.html
- AAAI 2021 eXplainable Recommender Systems (#1) http://www.inf.unibz.it/~rconfalonieri/aaai21/
- AAAI 2021 / NeurIPS 2020 Explaining Machine Learning Predictions: State-of-the-art, Challenges, and Opportunities (#2) - http://explainml-tutorial.github.io/ + video: https://www.youtube.com/watch?v=EbpU4p_0hes
- AAAI 2021 From Explainability to Model Quality and Back Again (#1)
- AAAI 2021 Tutorial On Explainable AI: From Theory to Motivation, Industrial Applications and Coding Practices (#3) - https://xaitutorial2019.github.io/ https://xaitutorial2020.github.io/
- IJCAI 2020 Tutorial on Logic-Enabled Verification and Explanation of ML Models (#1) - https://alexeyignatiev.github.io/ijcai20-tutorial/index.html
- ICIP 2018 / EMBC 2019 Interpretable Deep Learning: Towards Understanding & Explaining Deep Neural Networks (#2) - http://interpretable-ml.org/icip2018tutorial/ - http://interpretable-ml.org/embc2019tutorial/
- ICCV 2019 Tutorial on Interpretable Machine Learning for Computer Vision (#2) - https://interpretablevision.github.io/
- KDD 2019 Tutorial on Explainable AI in Industry (#1) - https://sites.google.com/view/kdd19-explainable-ai-tutorial

Workshop:

- BlackboxNLP 2020: Analyzing and interpreting neural networks for NLP (#3): https://blackboxnlp.github.io/
- IEEE VIS Workshop on Visualization for AI Explainability 2020 (#3) - https://visxai.io/
- ISWC 2020 Workshop on Semantic Explainability (#2) - http://www.semantic-explainability.com/
- IJCAI 2020 Workshop on Explainable Artificial Intelligence (#4) - https://sites.google.com/view/xai2020/home 55 paper submitted in 2019
- AAAI 2021 Workshop on Explainable Artificial Intelligence (#5 – follow-up of IJCAI serie)- https://sites.google.com/view/xaiworkshop/
- IJCAI 2019 Workshop on Optimisation and Explanation in AI (#1) - https://www.doc.ic.ac.uk/~kc2813/OXAI/
- SIGIR 2020 Workshop on Explainable Recommendation and Search (#3) https://ears2020.github.io
- ICAPS 2020 Workshop on Explainable Planning (#3)- https://kcl-planning.github.io/XAIP-Workshops/ICAPS_2019 23 papers submitted in 2019 https://icaps20subpages.icaps-conference.org/workshops/xaip/
- KDD 2019 Workshop on Explainable AI for fairness, accountability, and transparency (#1) – https://xai.kdd2019.a.intuit.com
- ICCV 2019 Workshop on Interpreting and Explaining Visual Artificial Intelligence Models (#1) - http://xai.unist.ac.kr/workshop/2019/
- NeurIPS 2019 Workshop on Challenges and Opportunities for AI in Financial Services: the Impact of Fairness, Explainability, Accuracy, and Privacy - https://sites.google.com/view/feap-ai4fin-2018/
- CD-MAKE 2021 – Workshop on Explainable AI (#4) - https://cd-make.net/make-explainable-ai/
- AAAI 2019 / CVPR 2019 Workshop on Network Interpretability for Deep Learning (#1 and #2) - http://networkinterpretability.org/ - https://explainai.net/
- IEEE FUZZ 2019 / Advances on eXplainable Artificial Intelligence (#2) - https://sites.google.com/view/xai-fuzzieee2019
- International Conference on NL Generation - Interactive Natural Language Technology for Explainable Artificial Intelligence (EU H2020 NL4XAI; #1) - https://sites.google.com/view/nl4xai2019/

Conference

- 2021 ACM Conference on Fairness, Accountability, and Transparency (ACM FAccT) (#4) https://facctconference.org/

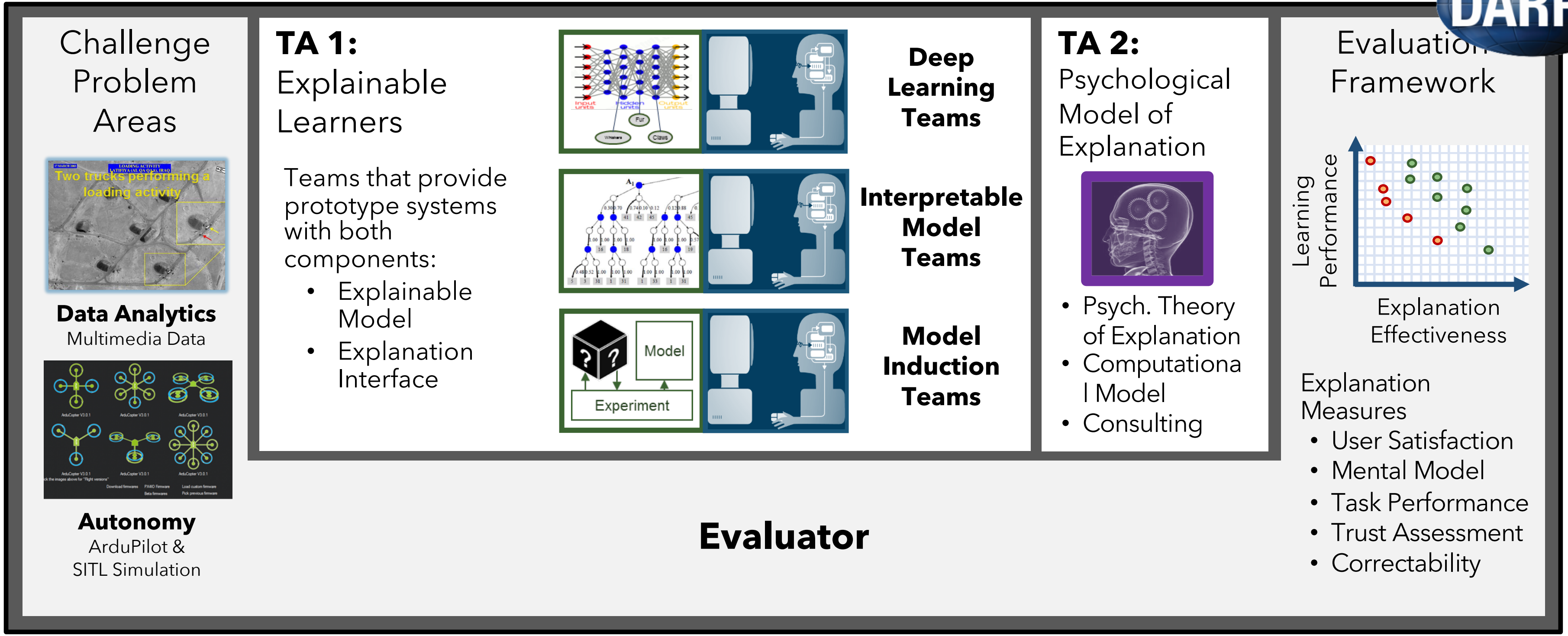
Challenge:

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

(Some) Software Resources

- Facebook Fairseq: https://github.com/pytorch/fairseq (to capture attention weights per input token… and much more)
- Saliency-based XAI: https://github.com/chihkuanyeh/saliency_evaluation + https://github.com/pair-code/saliency/blob/master/Examples.ipynb (Vanilla Gradients, Guided Backpropogation, Integrated Gradients, Occlusion)
- XAI Empirical studies: https://paperswithcode.com/paper/how-can-i-explain-this-to-you-an-empirical
- Facebook Captum - https://github.com/pytorch/captum
- IBM-MIT shared-interest https://github.com/aboggust/shared-interest
- Google-CMU Post-training Concept-based Explanation: https://github.com/chihkuanyeh/concept_exp
- Google-Stanford Automatic Concept-based Explanations: https://github.com/amiratag/ACE
- Google Testing with Concept Activation Vectors https://github.com/tensorflow/tcav
- DeepExplain: perturbation and gradient-based attribution methods for Deep Neural Networks interpretability. github.com/marcoancona/DeepExplain
- iNNvestigate: A toolbox to iNNvestigate neural networks' predictions. github.com/albermax/innvestigate
- SHAP: SHapley Additive exPlanations. github.com/slundberg/shap
- Microsoft Explainable Boosting Machines. https://github.com/Microsoft/interpret
- GANDissect: Pytorch-based tools for visualizing and understanding the neurons of a GAN. https://github.com/CSAILVision/GANDissect
- ELI5: A library for debugging/inspecting machine learning classifiers and explaining their predictions. github.com/TeamHG-Memex/eli5
- Skater: Python Library for Model Interpretation/Explanations. github.com/datascienceinc/Skater
- Yellowbrick: Visual analysis and diagnostic tools to facilitate machine learning model selection. github.com/DistrictDataLabs/yellowbrick
- Lucid: A collection of infrastructure and tools for research in neural network interpretability. github.com/tensorflow/lucid
- LIME: Agnostic Model Explainer. https://github.com/marcotcr/lime
- Sklearn_explain: model individual score explanation for an already trained scikit-learn model. https://github.com/antoinecarme/sklearn_explain
- Heatmapping: Prediction decomposition in terms of contributions of individual input variables
- Deep Learning Investigator: Investigation of Saliency, Deconvnet, GuidedBackprop and more. https://github.com/albermax/innvestigate
- Google PAIR What-if: Model comparison, counterfactual, individual similarity. https://pair-code.github.io/what-if-tool/
- Google tf-explain: https://tf-explain.readthedocs.io/en/latest/
- IBM AI Fairness: Set of fairness metrics for datasets and ML models, explanations for these metrics. https://github.com/IBM/aif360
- Blackbox auditing: Auditing Black-box Models for Indirect Influence. https://github.com/algofairness/BlackBoxAuditing
- Model describer: Basic statiscal metrics for explanation (visualisation for error, sensitivity). https://github.com/DataScienceSquad/model-describer
- *AXA Interpretability and Robustness: https://axa-rev-research.github.io/ (more on research resources – not much about tools)*

(Some) Initiatives: XAI in USA



TA1: Explainable Learners

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

TA2: Psychological Model of Explanation

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

(Some) Initiatives: XAI in Canada

- DEEL (Dependable Explainable Learning) Project 2019-2024
    - Research institutions

      ACRIAQ IVADO CRSNG NSERC

    - Industrial partners

      Bell Helicopter A Textron Company BOMBARDIER CAE THALES

    - Academic partners
        - Science and technology to develop new methods towards Trustable and Explainable AI

      IRT SAINT EXUPERY POLYTECHNIQUE MONTRÉAL UNIVERSITÉ LAVAL

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

Certificability
- Structural warranties
- Risk auto evaluation
- External audit

Explicability & Interpretability

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

(Some) Initiatives: XAI in EU

Conclusion

Why do we need XAI by the way?

- ***To empower*** individual against undesired effects of automated decision making
- ***To reveal*** and protect new vulnerabilities
- ***To implement*** the “right of explanation”
- ***To improve*** industrial standards for developing AI-powered products, increasing the trust of companies and consumers
- ***To help*** people make better decisions
- ***To align*** algorithms with human values
- ***To preserve*** (and expand) human autonomy
- **To scale and industrialize** AI

Conclusion

- Explainable AI is motivated by **real-world applications in AI – <u>Needs of Actionable XAI</u>**
- 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**
- **Need "Explainability by Design" when building AI products**

Open Research Questions

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

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.
- XAI as a methodology for debugging ML systems

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

Thanks! Questions?

- Feedback most welcome :-)
    - freddy.lecue@inria.fr (@freddy.lecue)
    - p.minervini@ucl.ac.uk,
    - riccardo.guidotti@unipi.it
- Tutorial website: https://xaitutorial2021.github.io
- To try Thales XAI Platform , please send an email to freddy.lecue@thalesgroup.com

THALES

UCL

INVENTEURS DU MONDE NUMÉRIQUE

IN SUPREMÆ DIGNITATIS 1343