# Explainable AI - XAI

***Watch the Semantic Gap!***

**Freddy Lecue (@freddylecue)**
http://www-sop.inria.fr/members/Freddy.Lecue/

Distinguished seminars on Explainable AI - by "XAI Science and technology for the eXplanation of AI decision making"

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

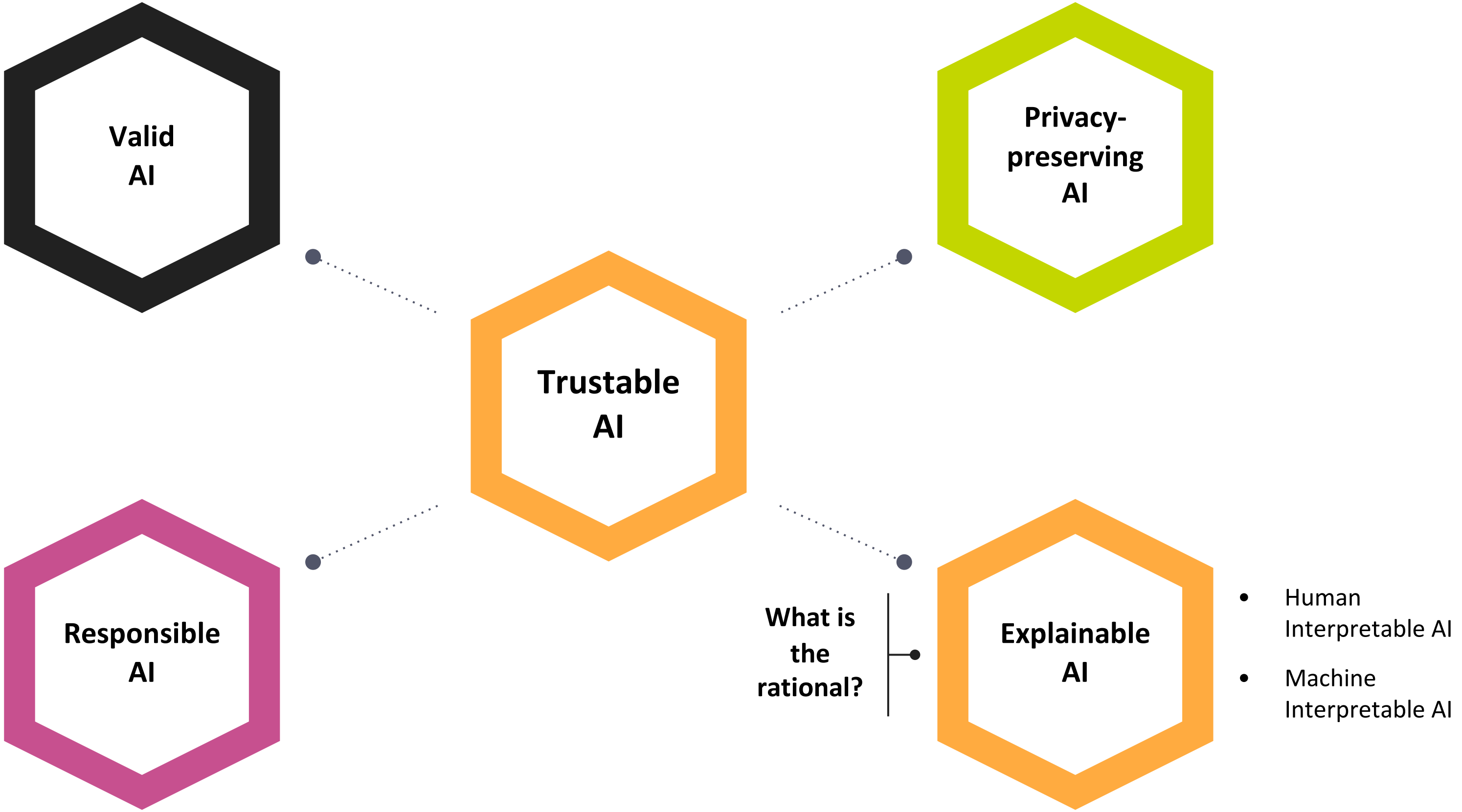
***July 13th, 2021***

https://tinyurl.com/9ahdbtm4

INVENTEURS DU MONDE NUMÉRIQUE

# Scope

# 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/

# 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


Opinion

OP-ED CONTRIBUTOR

When a Computer Program Keeps You in Jail

By Rebecca Wexler

June 13, 2017

# … but not only Critical Systems (2)

## Finance:

- Credit scoring, loan approval
- Insurance quotes

FICO COMMUNITY


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

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

**Insurance: Robots learn the business of covering risk**

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

Save

**Oliver Ralph** MAY 16, 2017 24

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

# … 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

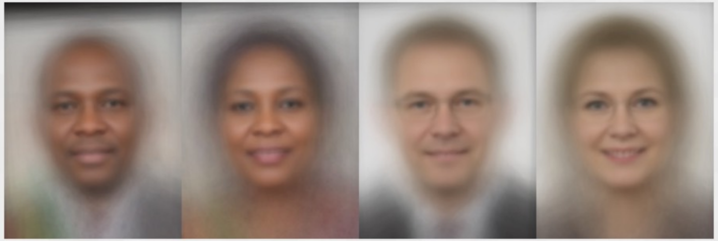
Marc Sturm
NewYork-Presbyterian Hospital
mas9161@nyp.org

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

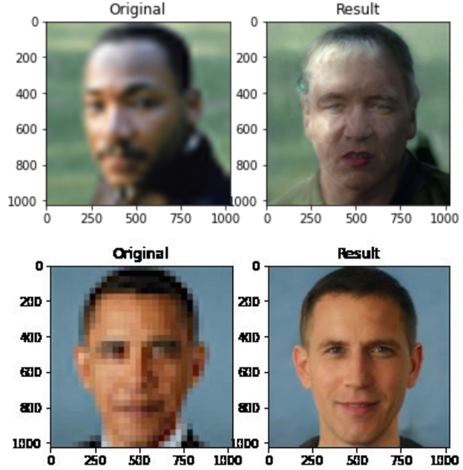
| Gender Classifier | Darker Male | Darker Female | Lighter Male | Lighter Female | Largest Gap |
|---|---|---|---|---|---|
| Microsoft | 94.0% | 79.2% | 100% | 98.3% | 20.8% |
| FACE++ | 99.3% | 65.5% | 99.2% | 94.0% | 33.8% |
| IBM | 88.0% | 65.3% | 99.7% | 92.9% | 34.4% |



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/

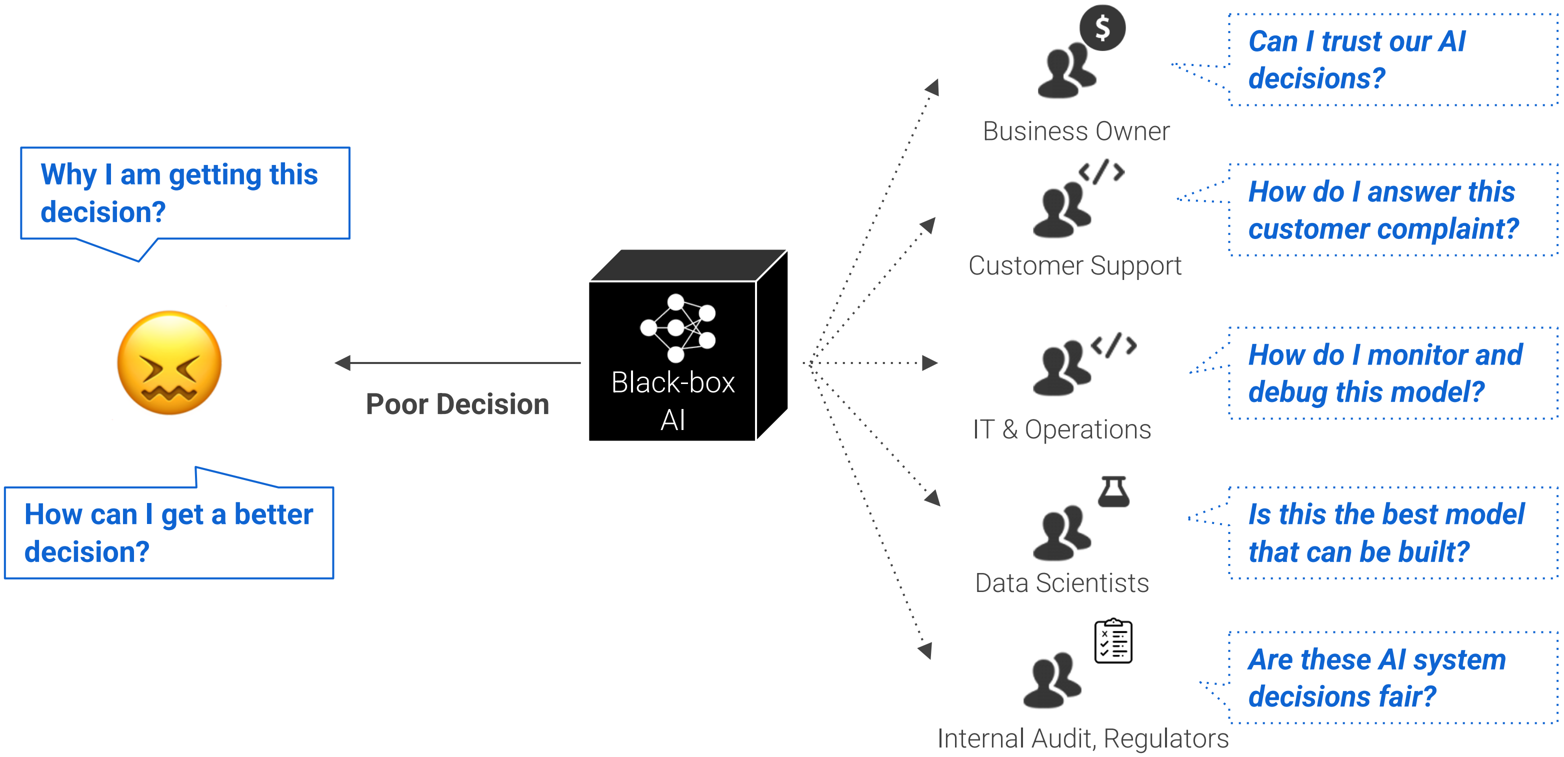

https://www.cbsnews.com/news/apple-credit-card-goldman-sachs-disputes-claims-that-apple-card-is-sexist/




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

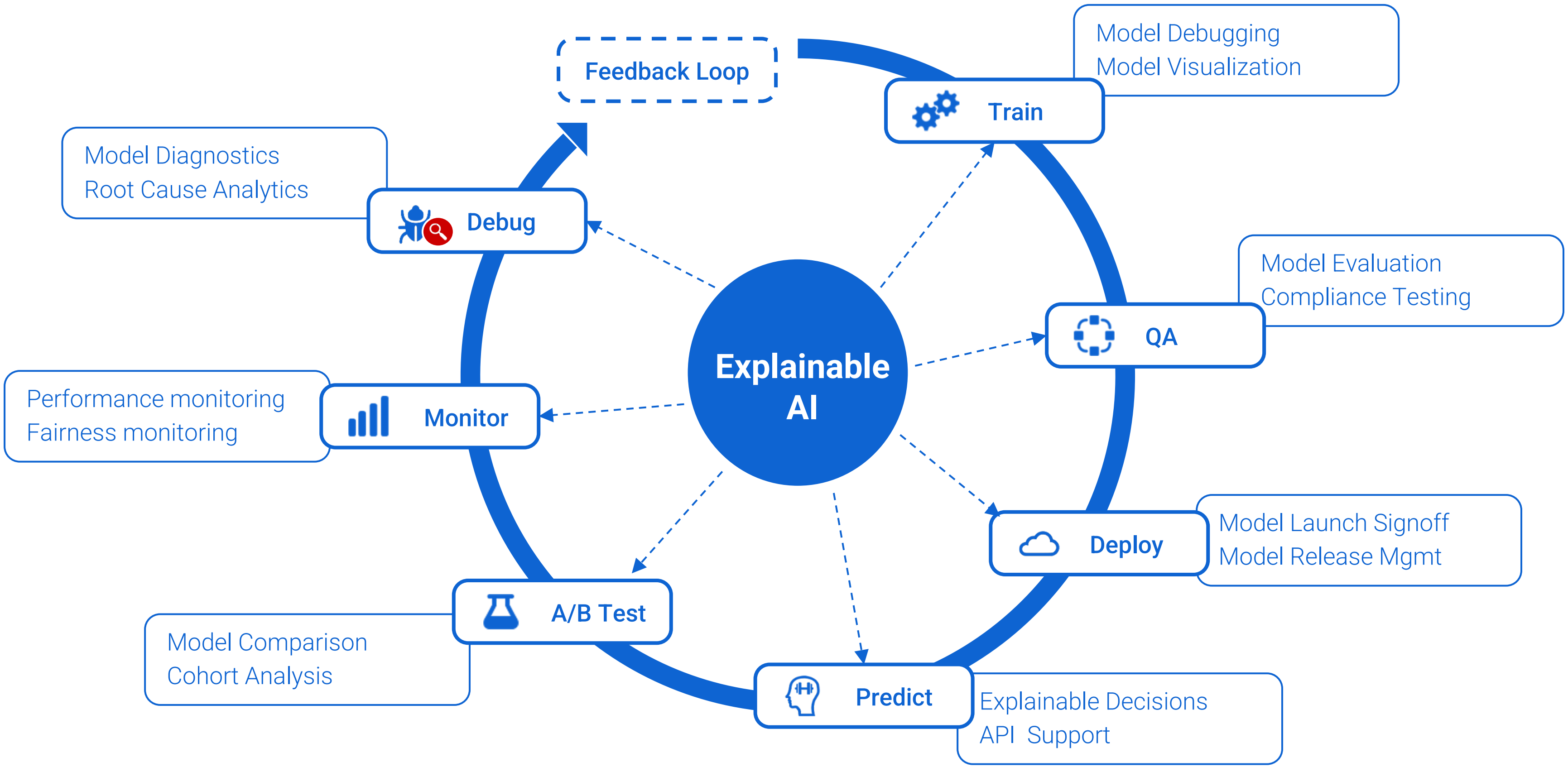
# Explanation - In a Nutshell

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



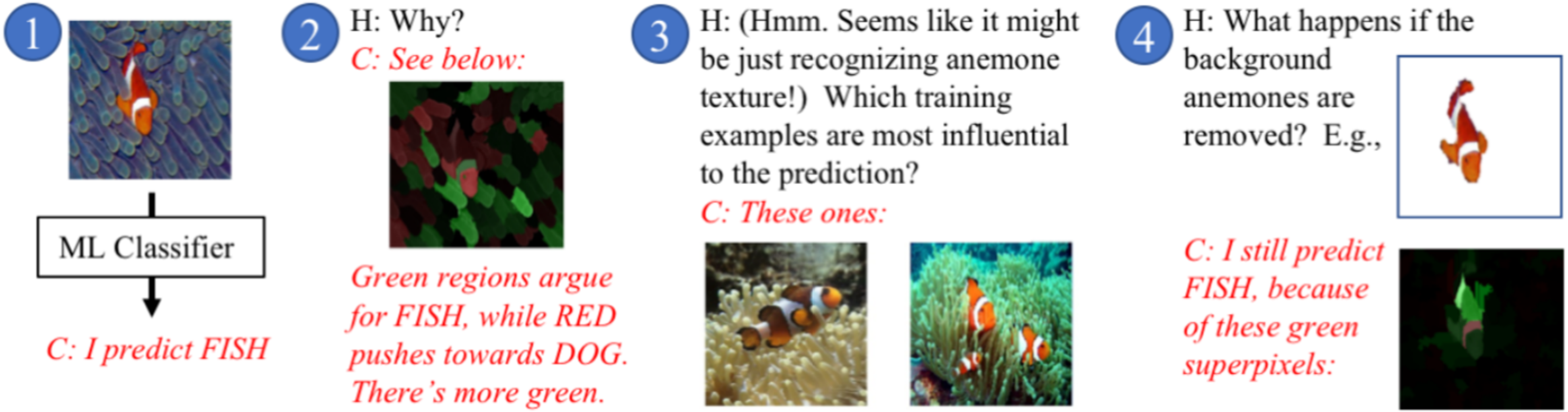

Credit: Lecue et al., Tutorial on XAI. AAAI 2020. https://xaitutorial2020.github.io/

# Explainability by Design for AI products




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

# 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).

# Evaluation - 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?

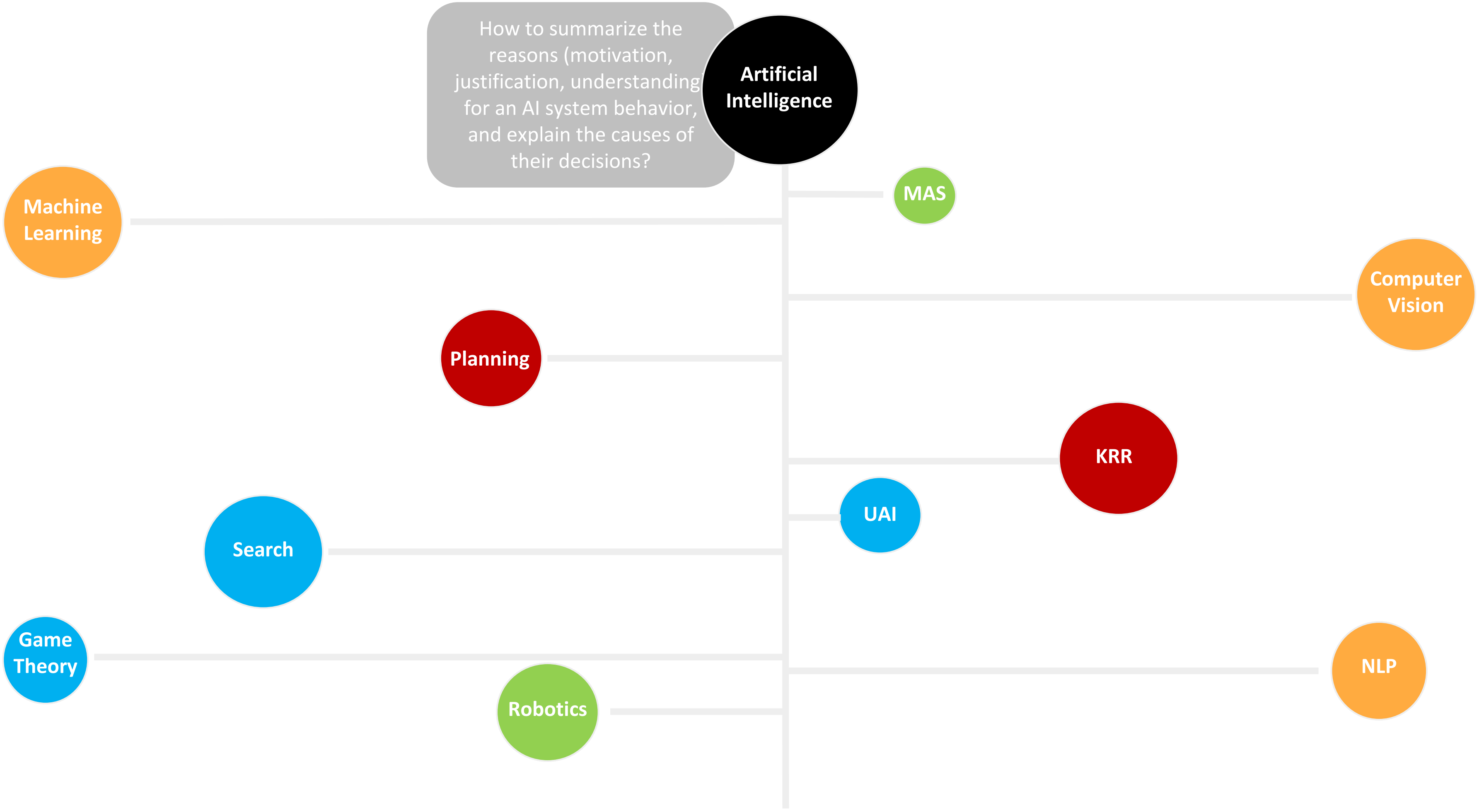
Source: Accenture Point of View. Understanding Machines: Explainable AI. Freddy Lecue, Dadong Wan

# Part II

## Explanation in AI (Focus Machine Learning)

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

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

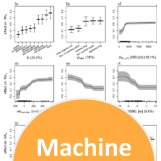


# XAI: One Objective, Many 'AI's, Many Definitions, Many Approaches

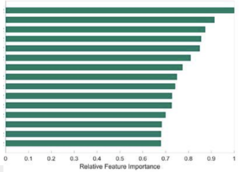
How to summarize the reasons (motivation, justification, understanding) for an AI system behavior, and explain the causes of their decisions?

**Artificial Intelligence**

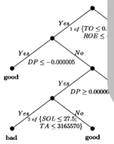
Dependency Plot



Feature Importance




Surrogate Model



**Machine Learning**

Which features are responsible of classification?

**MAS**

**Computer Vision**

**Planning**

**KRR**

**UAI**

**Search**

**Game Theory**

**NLP**

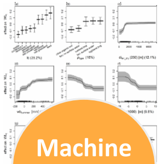
**Robotics**

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

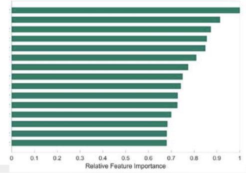
How to summarize the reasons (motivation, justification, understanding) for an AI system behavior, and explain the causes of their decisions?

Artificial Intelligence

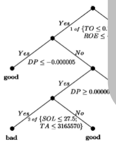
Dependency Plot



Feature Importance



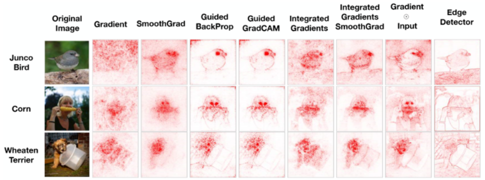

Surrogate Model




Machine Learning

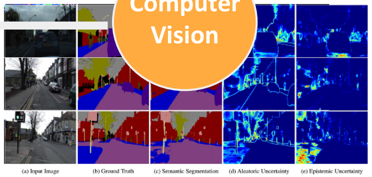
Which features are responsible of classification?

Saliency Map




Which complex features are responsible of classification?

Computer Vision




Uncertainty Map

MAS

Planning

KRR

UAI

Search

Game Theory

Robotics

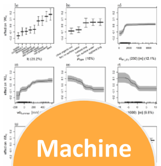
NLP

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

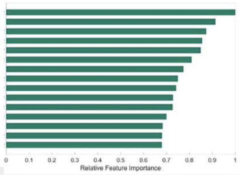
Artificial Intelligence

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

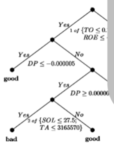
Dependency Plot



Feature Importance




Surrogate Model




Machine Learning

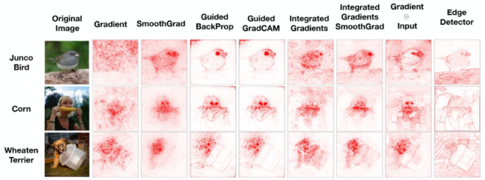
Which features are responsible of classification?

Strategy Summarization


MAS

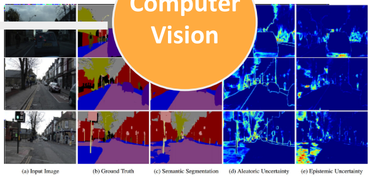
- Which agent strategy & plan ?
- Which player contributes most?
- Why such a conversational flow?

Saliency Map




Which complex features are responsible of classification?

Computer Vision




Uncertainty Map

Planning

KRR

UAI

Search

Game Theory

Robotics

NLP

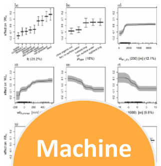
# XAI: One Objective, Many 'AI's, Many Definitions, Many Approaches

**Artificial Intelligence**

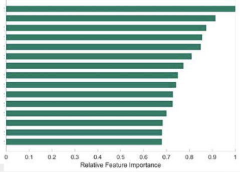
How to summarize the reasons (motivation, justification, understanding) for an AI system behavior, and explain the causes of their decisions?

**Machine Learning**

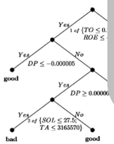
Dependency Plot



Feature Importance



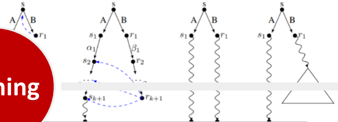

Surrogate Model




Which features are responsible of classification?

**Planning**

Plan Refinement



Which actions are responsible of a plan?

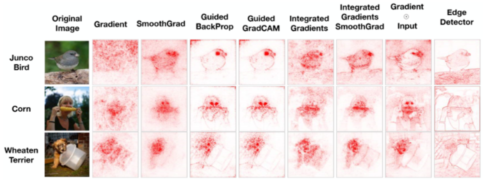
**MAS**

Strategy Summarization

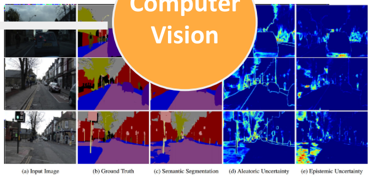

- Which agent strategy & plan ?
- Which player contributes most?
- Why such a conversational flow?

**Computer Vision**

Saliency Map




Which complex features are responsible of classification?




Uncertainty Map

**KRR**

**UAI**

**Search**

**Game Theory**

**Robotics**

**NLP**

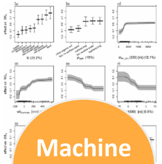
# XAI: One Objective, Many ‘AI’s, Many Definitions, Many Approaches

**Artificial Intelligence**

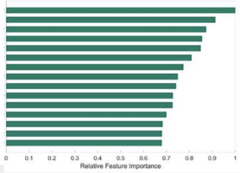
How to summarize the reasons (motivation, justification, understanding) for an AI system behavior, and explain the causes of their decisions?

**Machine Learning**

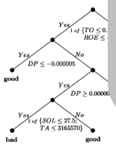
Dependency Plot



Feature Importance



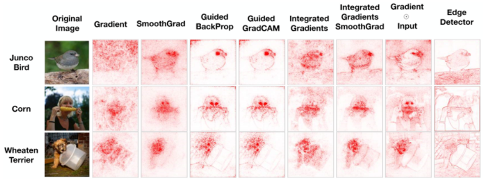

Surrogate Model



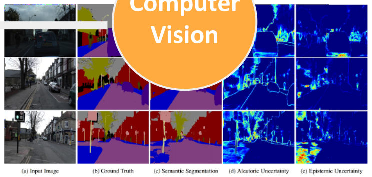
Which features are responsible of classification?

**Computer Vision**

Saliency Map




Which complex features are responsible of classification?




Uncertainty Map

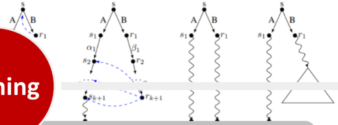
**MAS**

Strategy Summarization

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

**Planning**

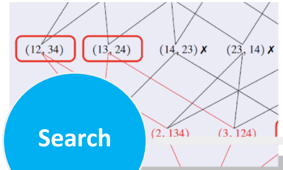
Plan Refinement



Which actions are responsible of a plan?

**Search**

Conflicts Resolution



Which constraints can be relaxed?

**KRR**

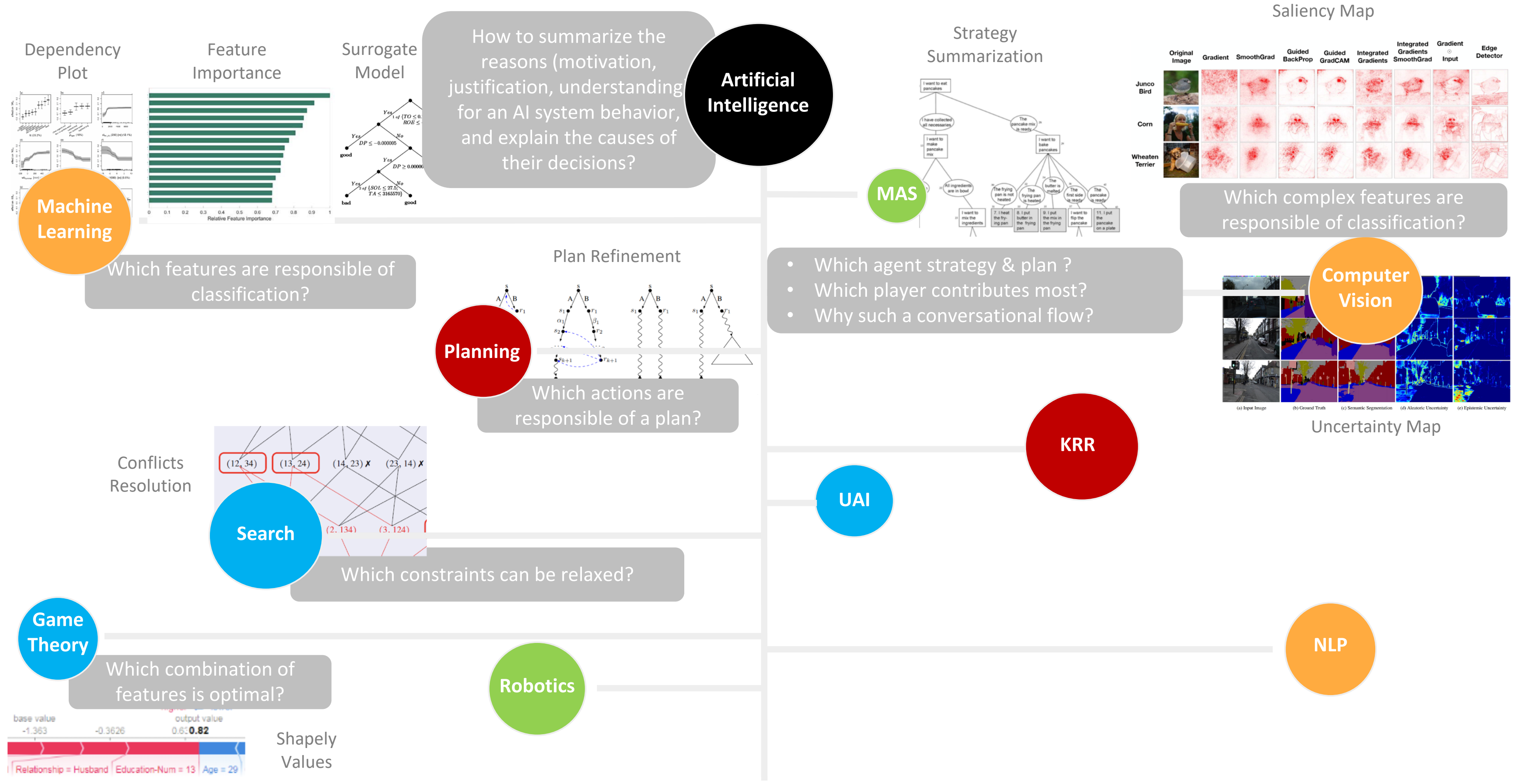
**UAI**

**Game Theory**

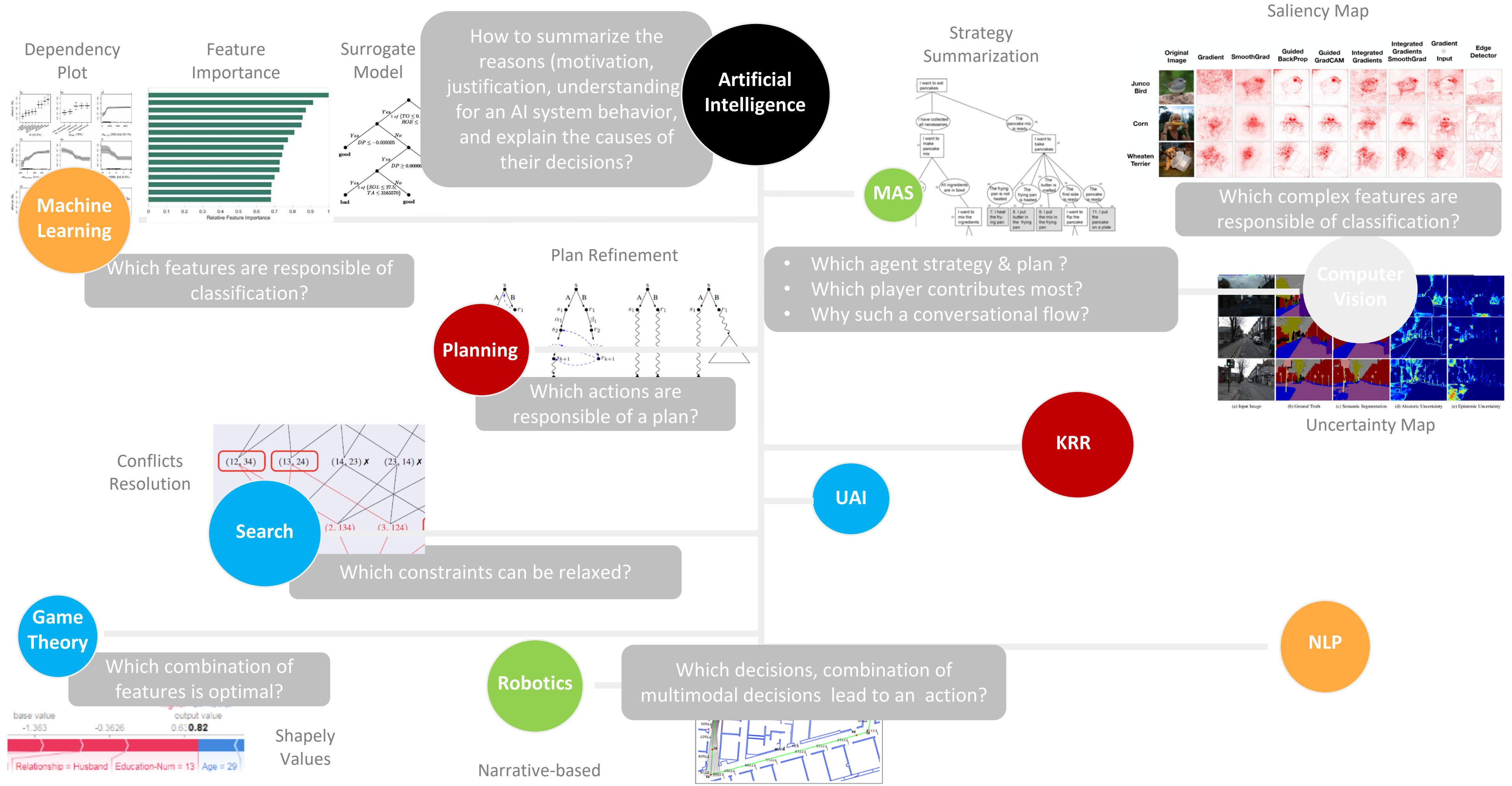
**Robotics**

**NLP**

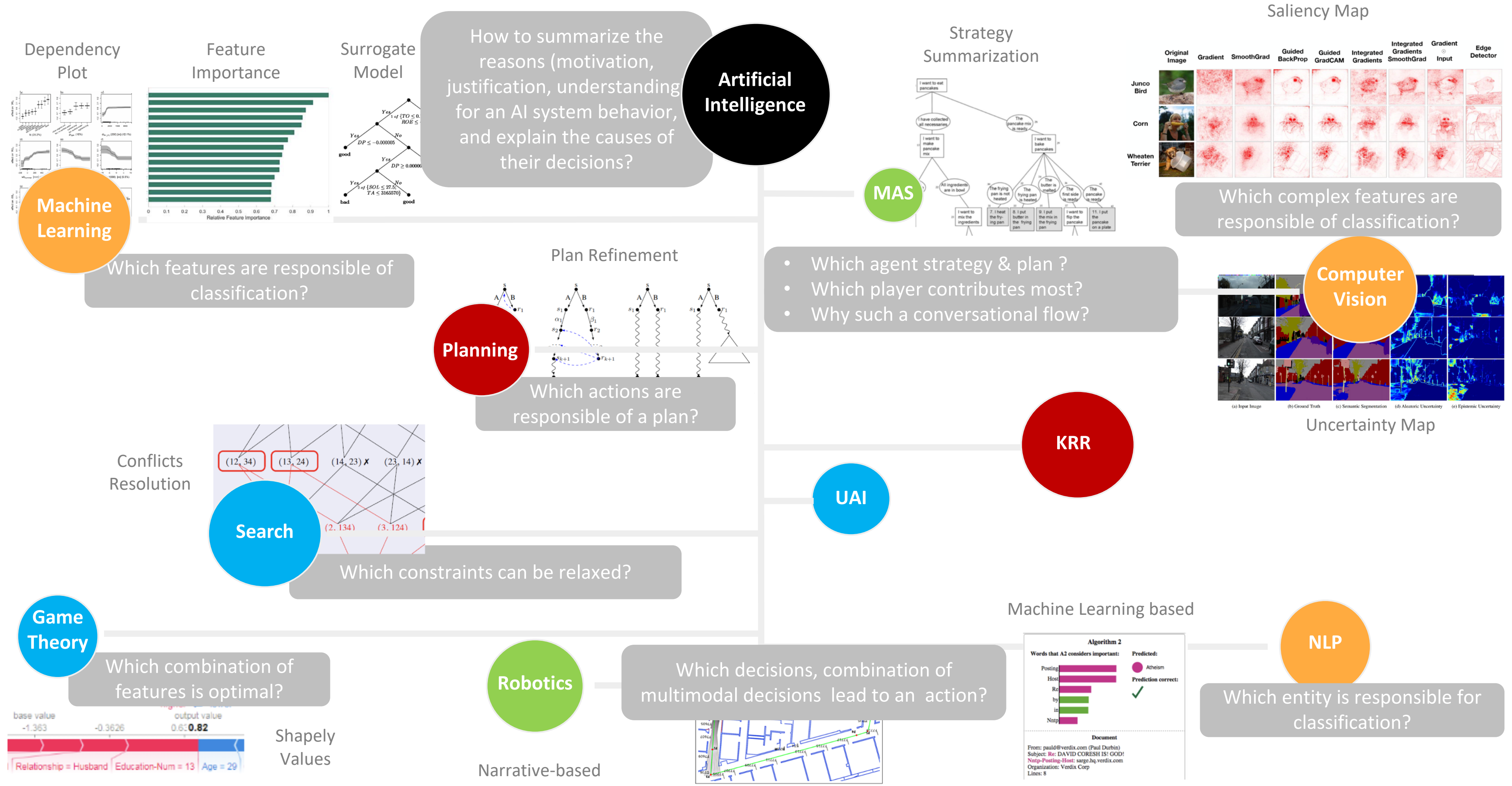
# XAI: One Objective, Many 'AI's, Many Definitions, Many Approaches



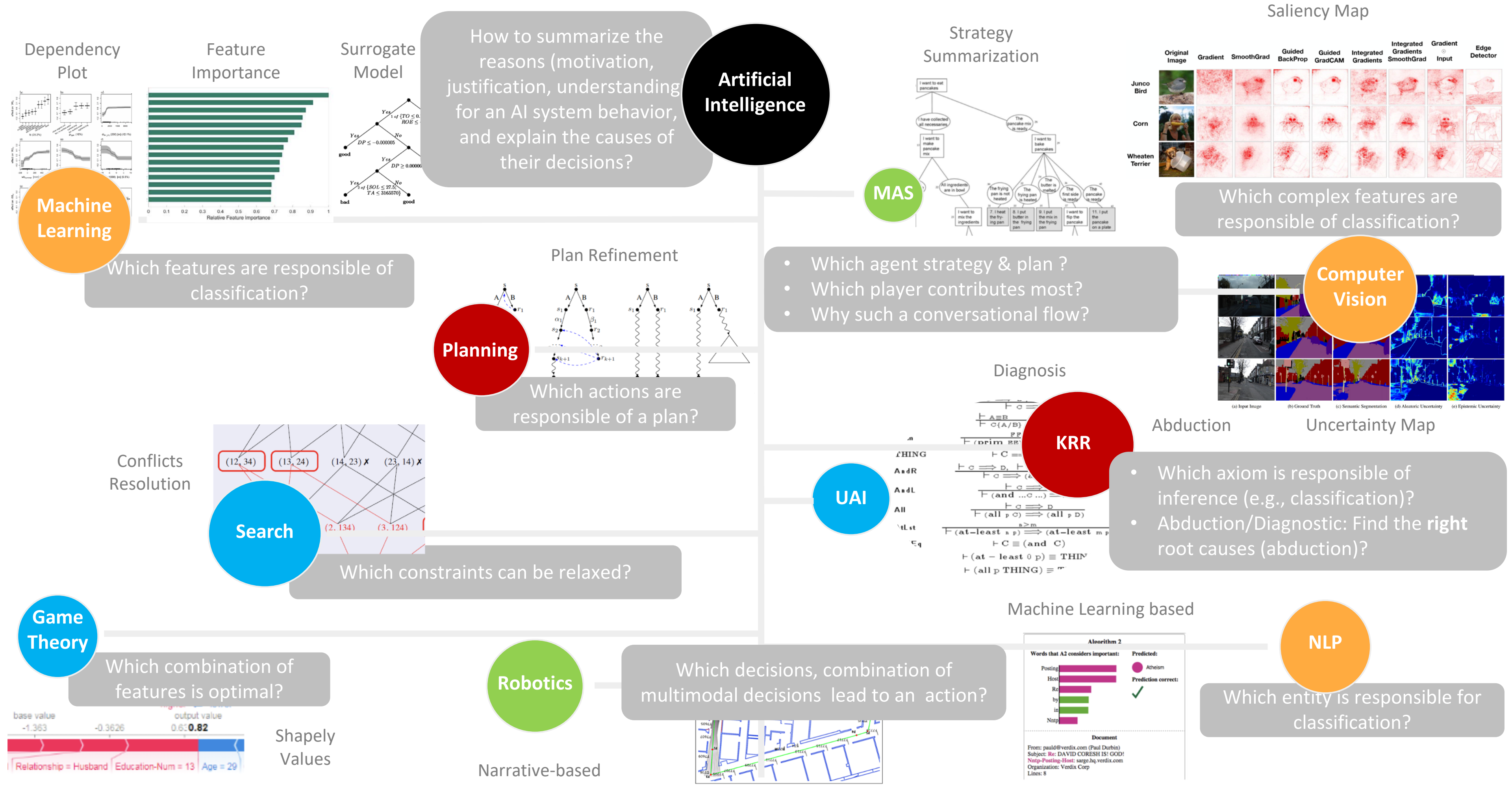
# XAI: One Objective, Many ‘AI’s, Many Definitions, Many Approaches



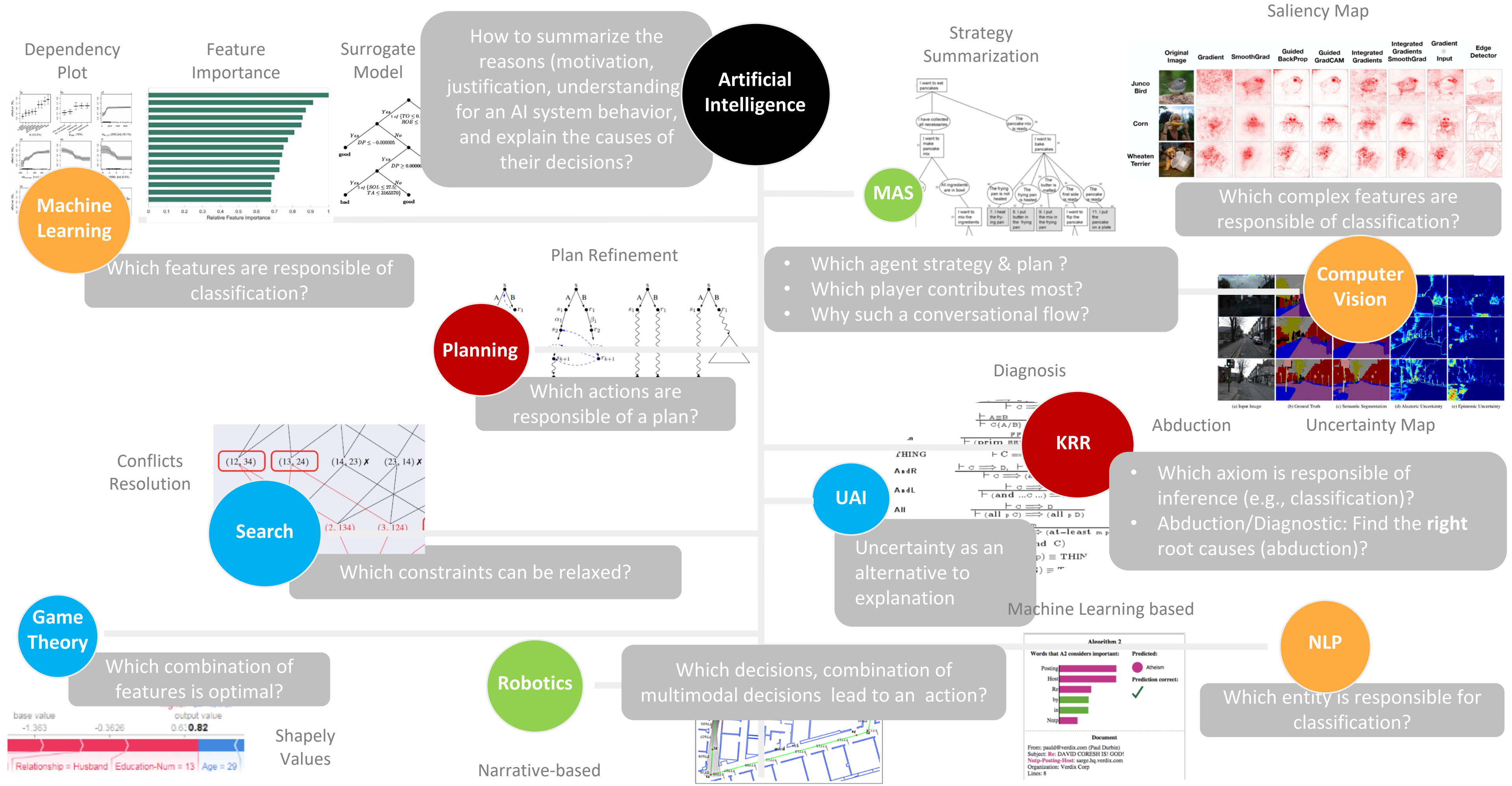
# XAI: One Objective, Many ‘AI’s, Many Definitions, Many Approaches



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



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

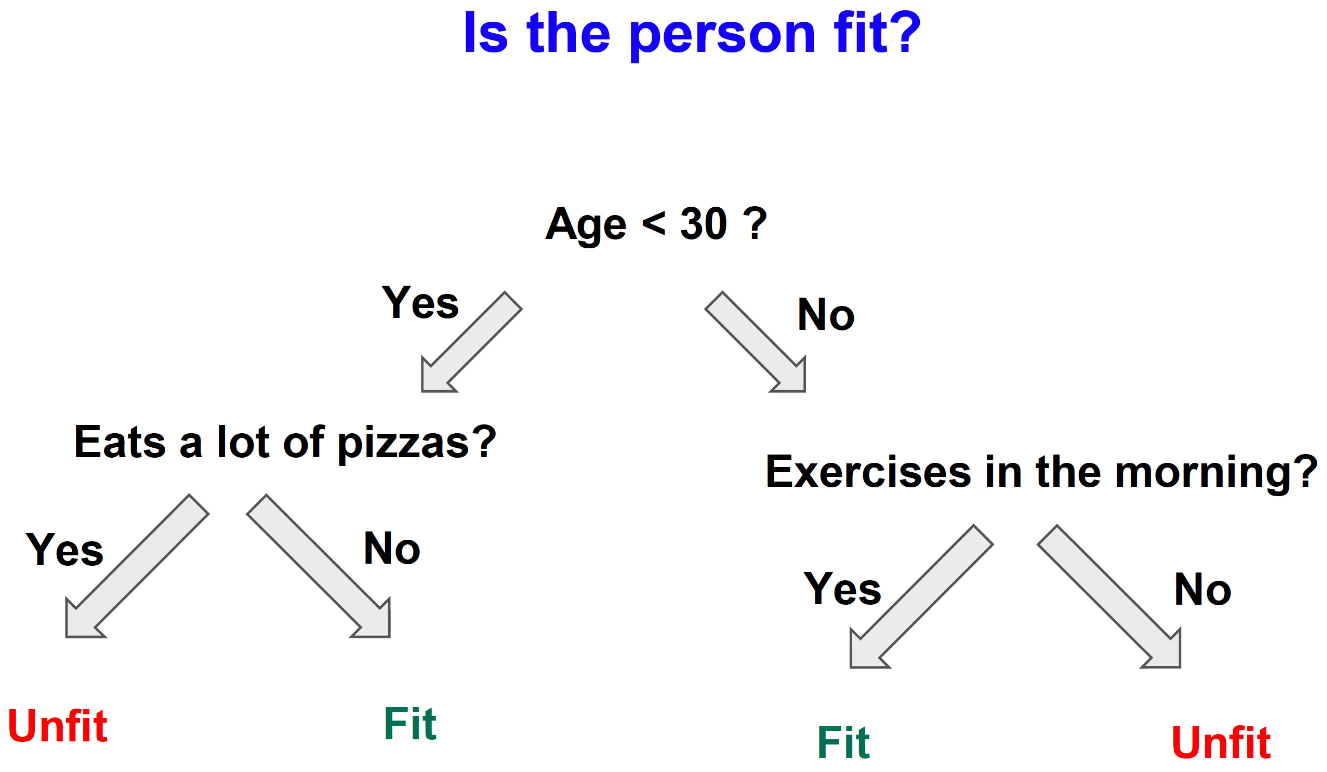


# Overview of Explanation in Machine Learning (1)

- Many tools already available from early-days Machine Learning

**Interpretable Models**:

- Decision Trees



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

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

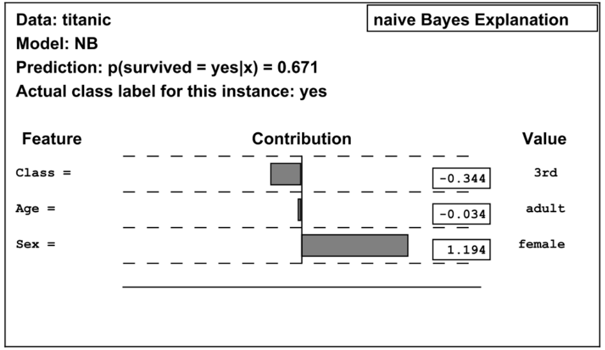
Accurate Intelligible Models with Pairwise Interactions. Lou, Caruana, Gehrke and Hooker. KDD 2013

# Overview of Explanation in Machine Learning (1)

- 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



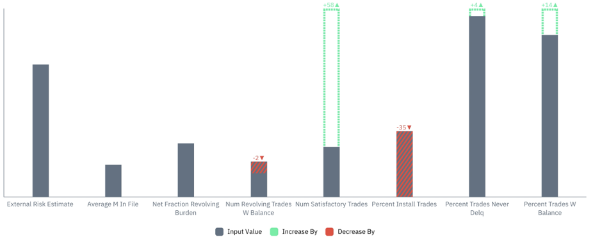

**Naive Bayes model**

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

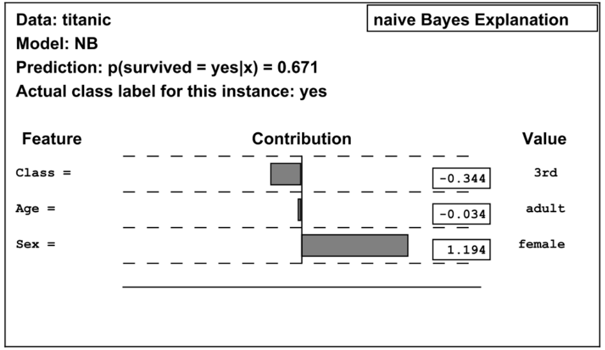



**Counterfactual What-if**

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**

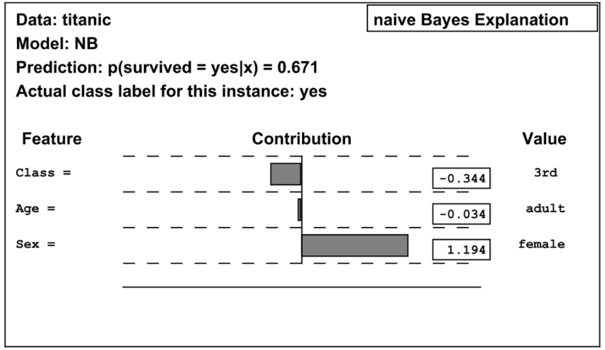
Igor Kononenko. Machine learning for medical diagnosis: history, state of the art and perspective. Artificial Intelligence in Medicine, 23:89–109, 2001.

# Overview of Explanation in Machine Learning (1)

- Many tools already available from early-days Machine Learning

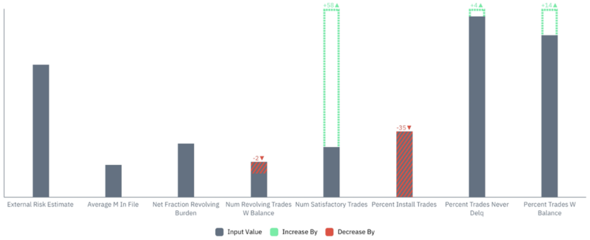
**Interpretable Models**:

- 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.

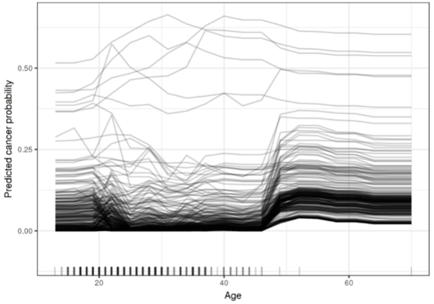


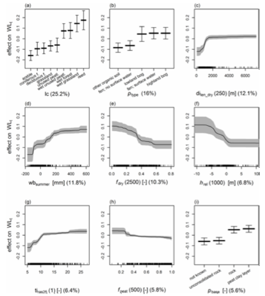

**Counterfactual What-if**

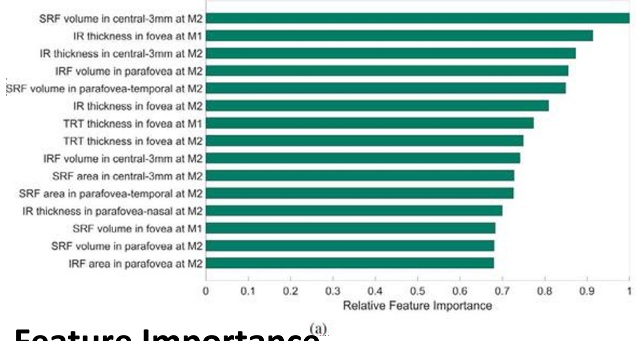
Brent D. Mittelstadt, Chris Russell, Sandra Wachter: Explaining Explanations in AI. FAT 2019: 279-288

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/



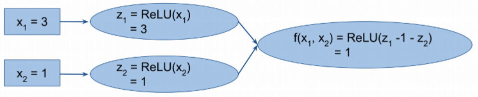






- **Feature Importance**
- **Partial Dependence Plot**
- **Individual Conditional Expectation**
- **Sensitivity Analysis**

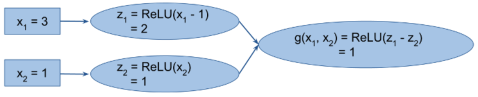
# Overview of Explanation in Machine Learning (2)

- 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 | $x_1 = 1.5,\ x_2 = -0.5$ |



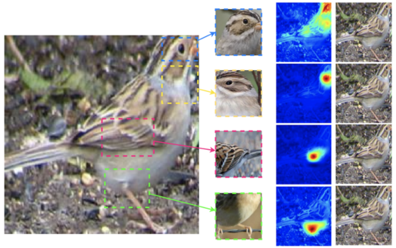

Network $g(x_1, x_2)$
Attributions at $x_1 = 3, x_2 = 1$

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

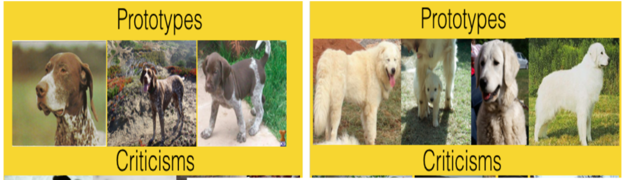
## Attribution for Deep Network (Integrated gradient-based)

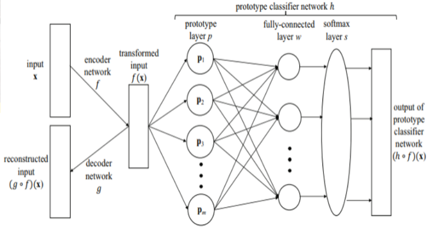
Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In ICML, pp. 3319–3328, 2017.

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)

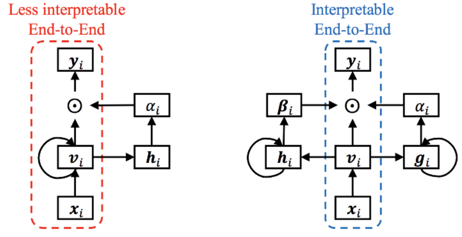



## Example-based / Prototype

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

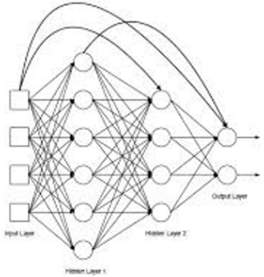


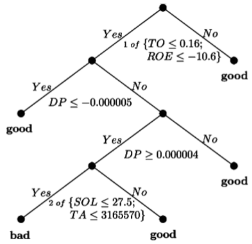

(a) Standard attention model (b) RETAIN model

## 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

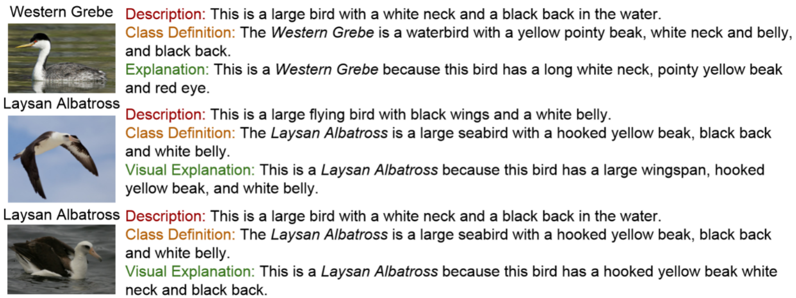
# Overview of Explanation in Machine Learning (3)

- Focus: Artificial Neural Network



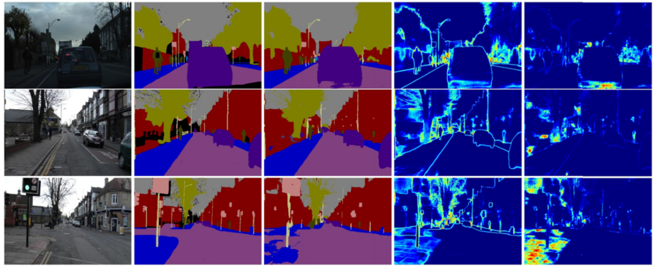
**Interpretable Units**

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




**Visual Explanation**

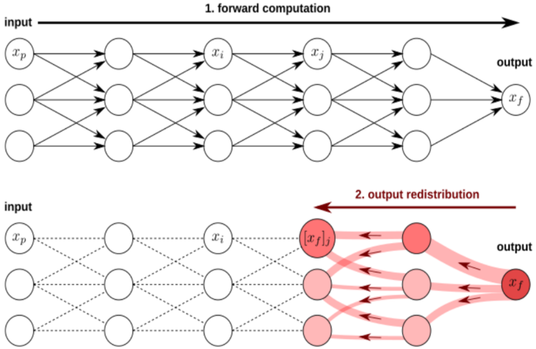
Lisa Anne Hendricks, Zeynep Akata, Marcus Rohrbach, Jeff Donahue, Bernt Schiele, Trevor Darrell: Generating Visual Explanations. ECCV (4) 2016: 3-19

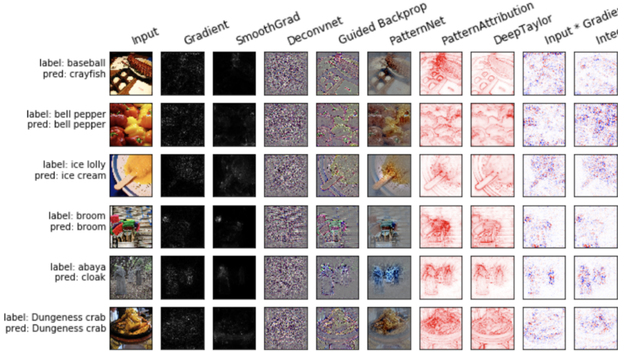


(a) Input Image (b) Ground Truth (c) Semantic Segmentation (d) Aleatoric Uncertainty (e) Epistemic Uncertainty

**Uncertainty Map**

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



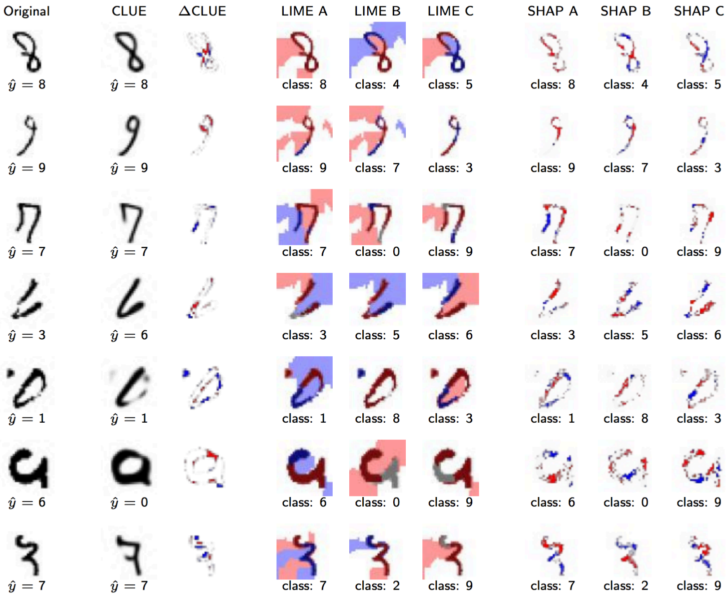



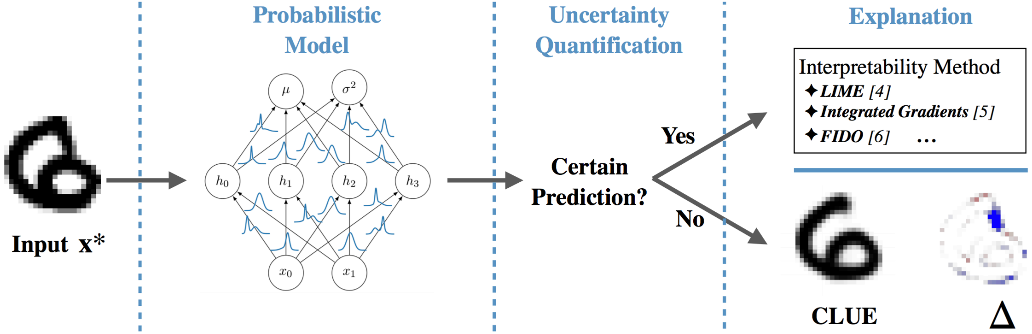

**Saliency Map / Features Attribution-based**

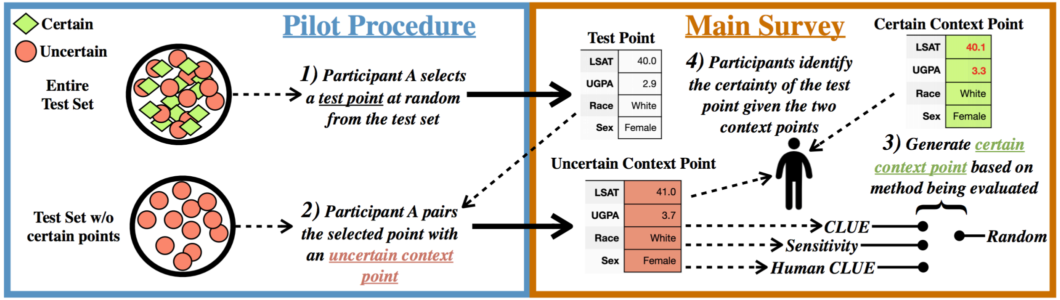
Julius Adebayo, Justin Gilmer, Michael Muelly, Ian J. Goodfellow, Moritz Hardt, Been Kim: Sanity Checks for Saliency Maps. NeurIPS 2018: 9525-9536

# 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

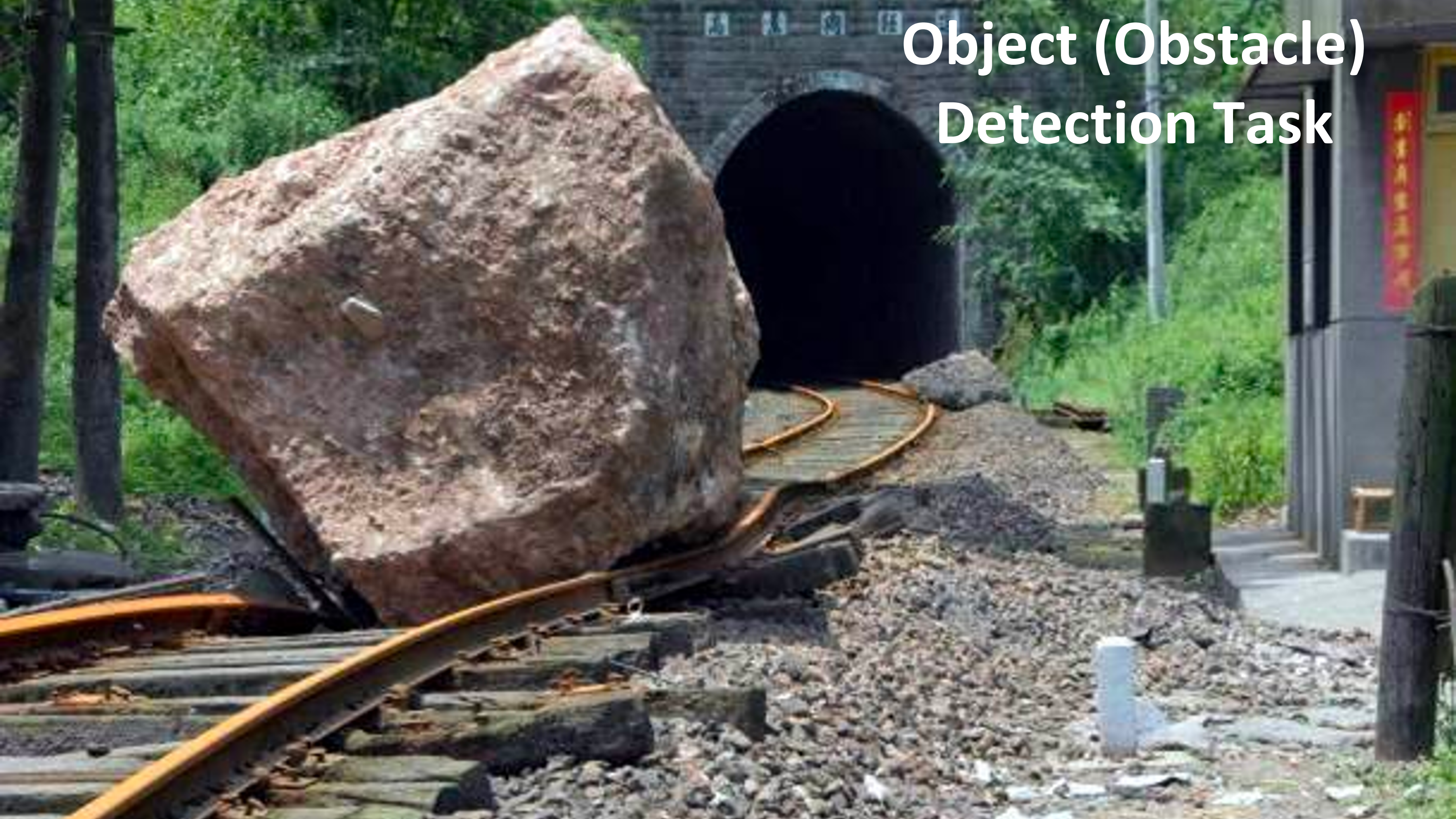
# Part III

## Watch the Semantic Gap

# How Does
# it
# Work
# in Practice?

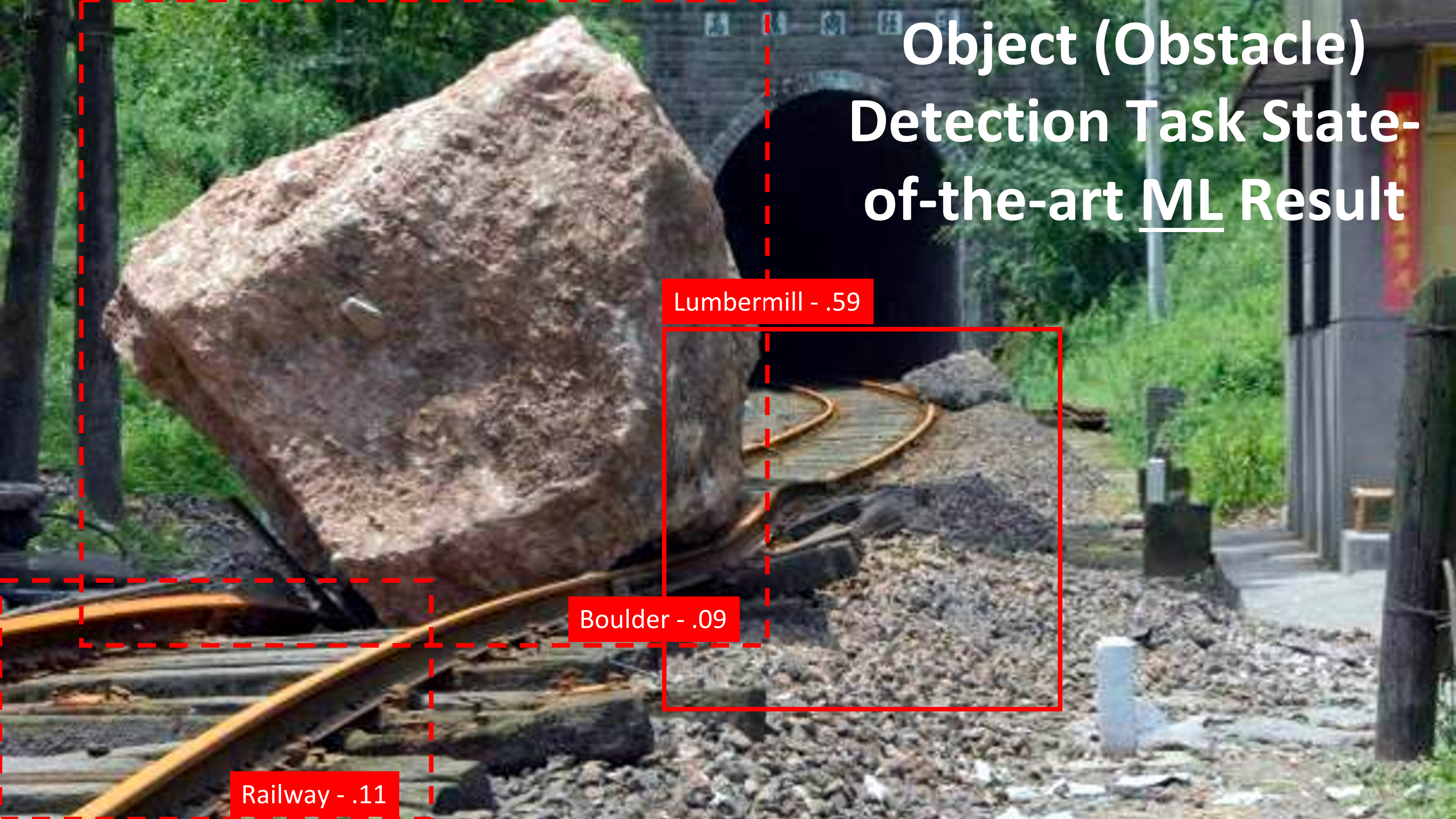
# State of the Art Machine Learning Applied to Critical Systems

# Object (Obstacle) Detection Task



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

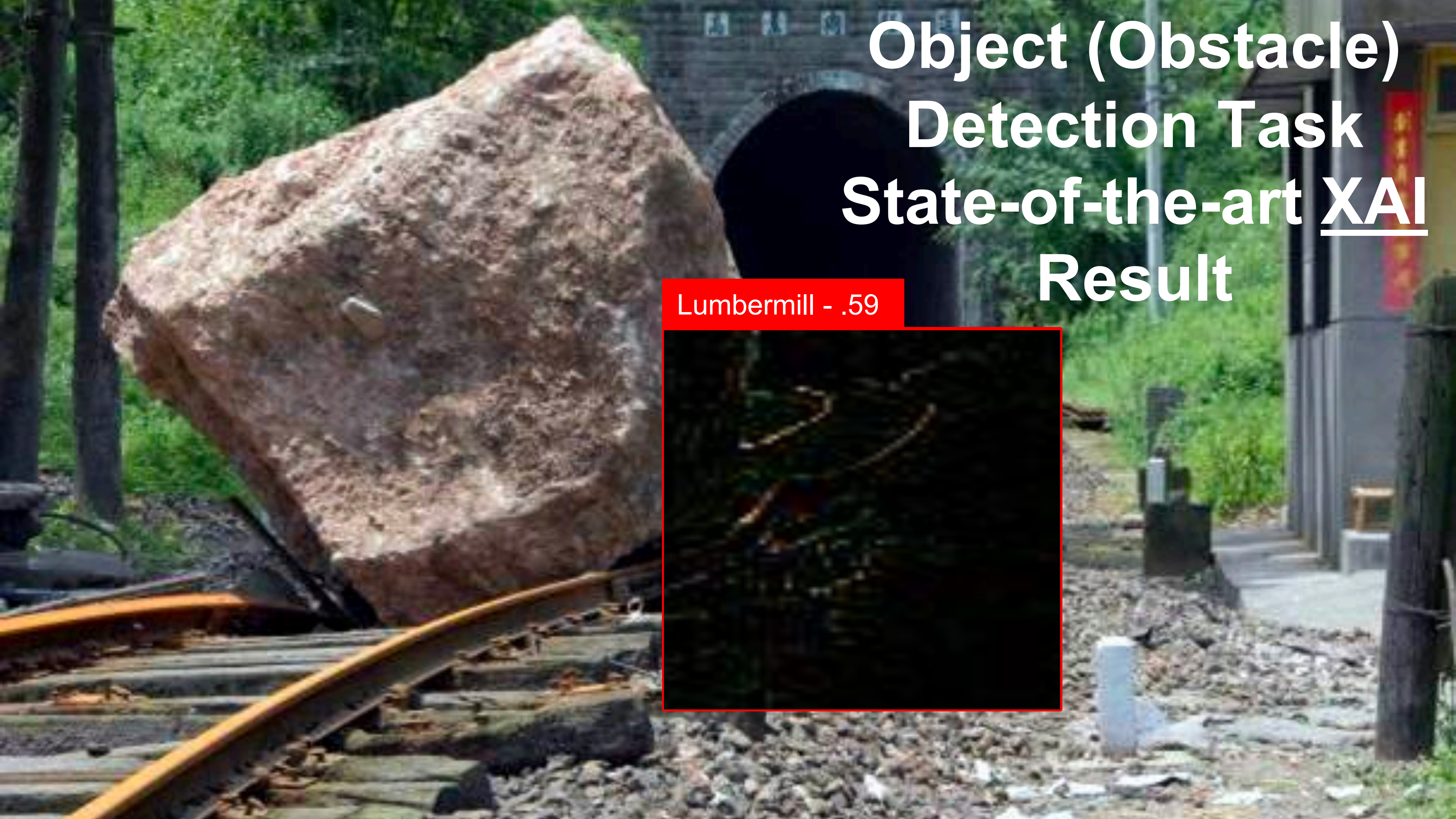
# Object (Obstacle) Detection Task State-of-the-art ML Result



# State of the Art
# XAI
# Applied to Critical Systems


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

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




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

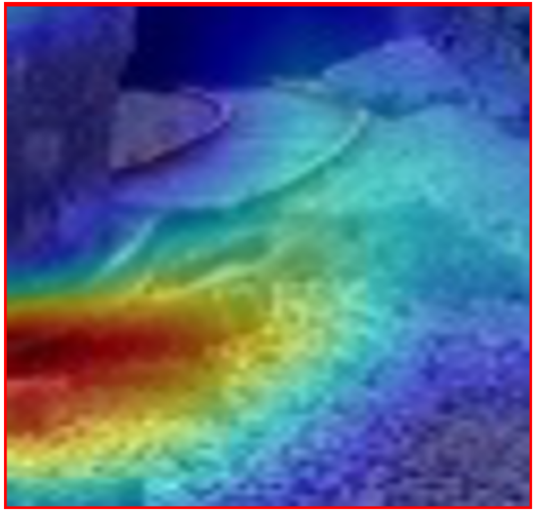
Unfortunately, this is of NO use for a human behind the system

# Let’s stay back

## Why this Explanation?
## (meta explanation)

After Human Reasoning…

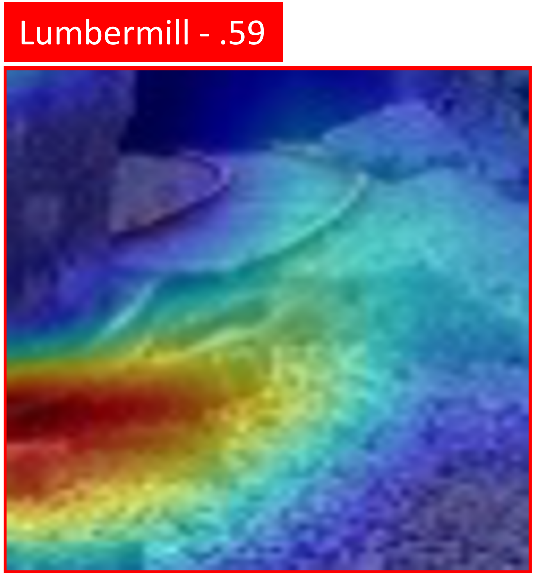
Lumbermill - .59

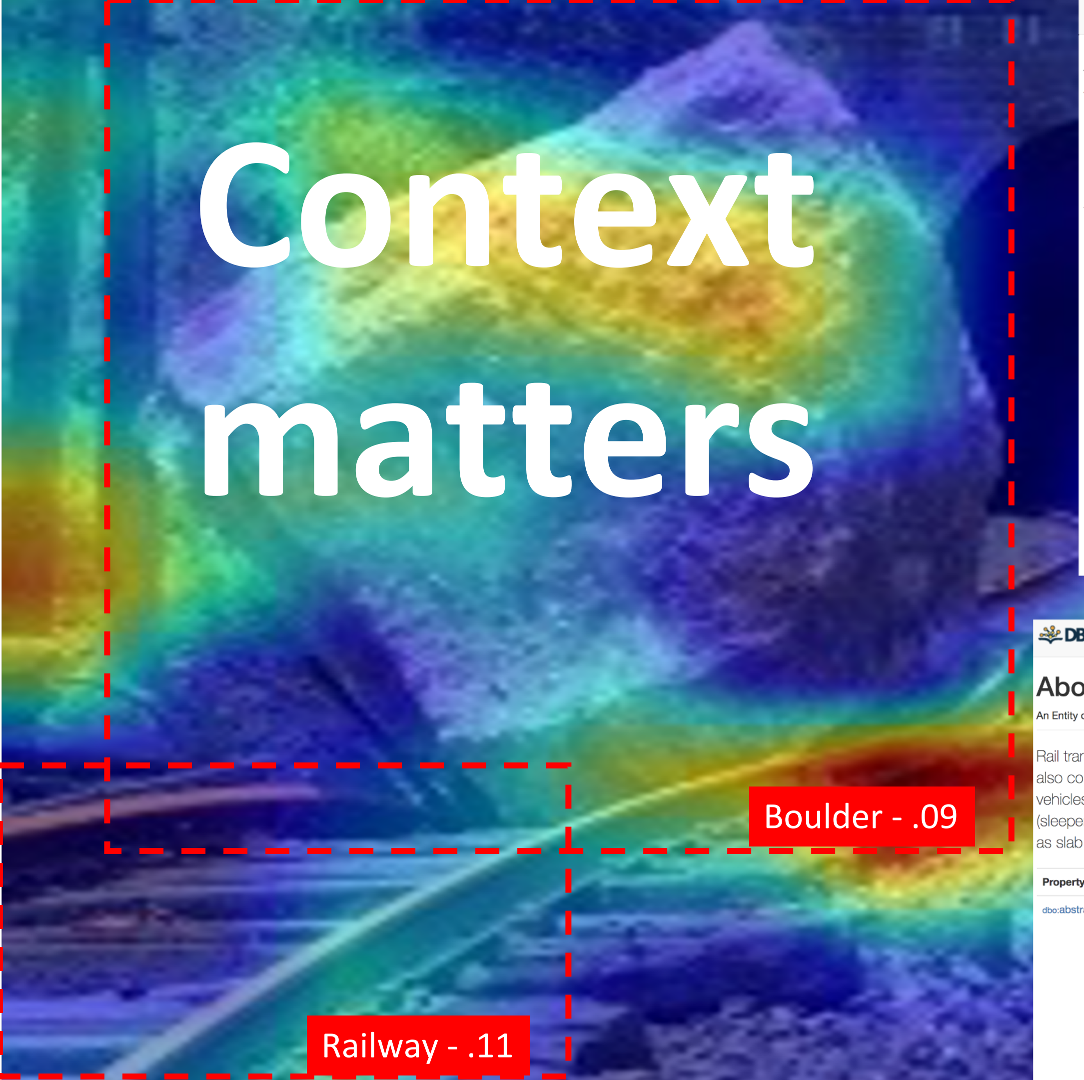


DBpedia | Browse using | Formats | Faceted Browser | Sparql Endpoint

| | |
|---|---|
| dbo:wikiPageID | 352327 (xsd:integer) |
| dbo:wikiPageRevisionID | 734430894 (xsd:integer) |
| dct:subject | dbc:Sawmills<br>dbc:Saws<br>dbc:Ancient_Roman_technology<br>dbc:Timber_preparation<br>dbc:Timber_industry |
| http://purl.org/linguistics/gold/hypernym | dbr:Facility |
| rdf:type | owl:Thing<br>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<br>dbpedia-cs:Sawmill<br>dbpedia-de:Sawmill<br>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<br>• dbc:Rocks |

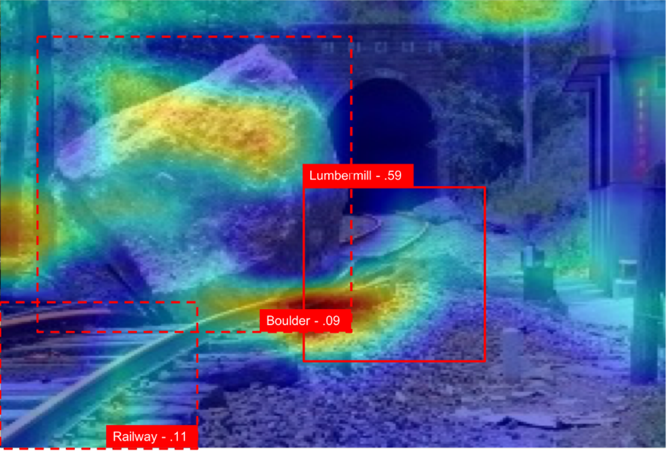
DBpedia Browse using Formats Faceted Browser Sparql Endpoint

# About: Rail transport

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

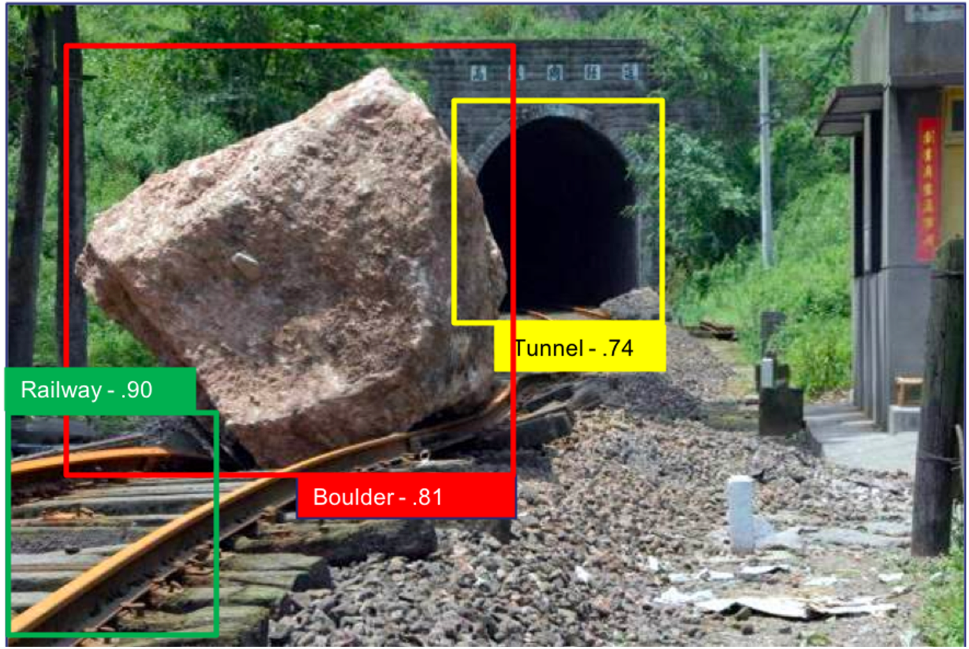
Rail transport is a means of conveyance of passengers and goods on wheeled vehicles running on rails, also known as tracks. It is also commonly referred to as train transport. In contrast to road transport, where vehicles run on a prepared flat surface, rail vehicles (rolling stock) are directionally 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

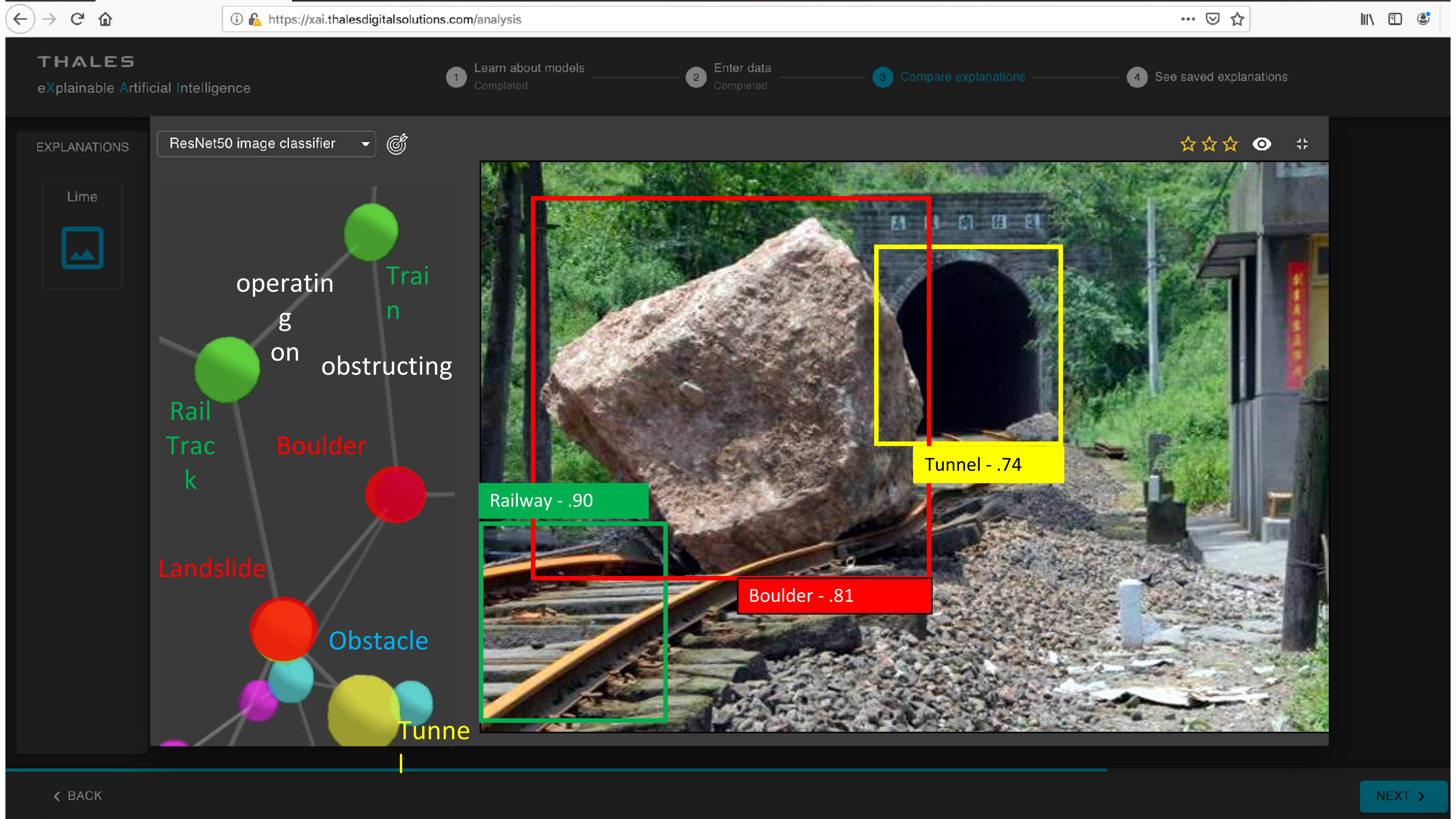
**Transitioning**



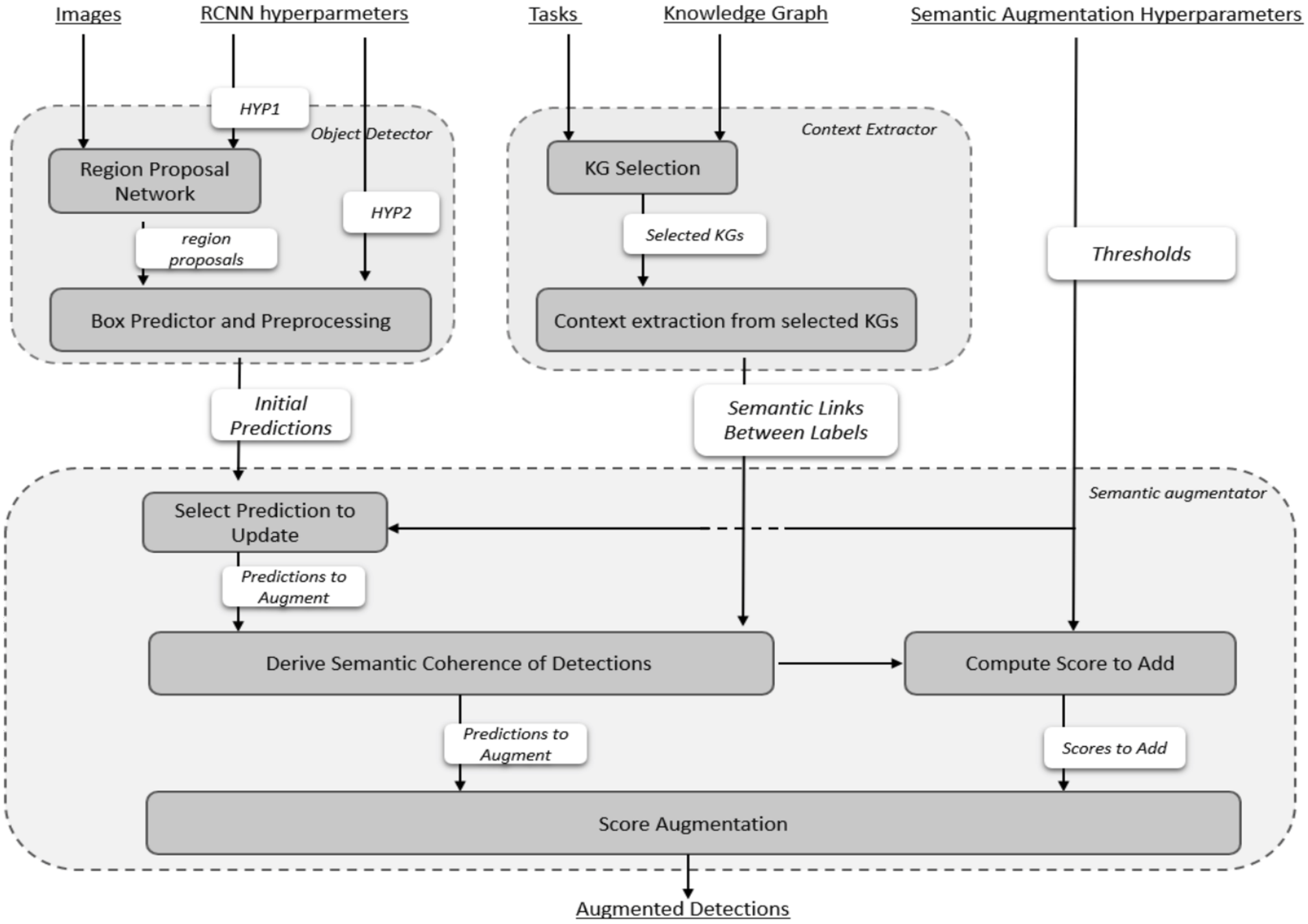

This is an **Obstacle: Boulder** obstructing the train: XG142-R on **Rail_Track** from City: Cannes to City: Marseille at **Location: Tunnel VIX** due to **Landslide**

# XAI Thales Platform

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


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
operatin
g
on
Trai
n
obstructing
Rail
Trac
k
Boulder
Landslide
Obstacle
Tunne
l
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

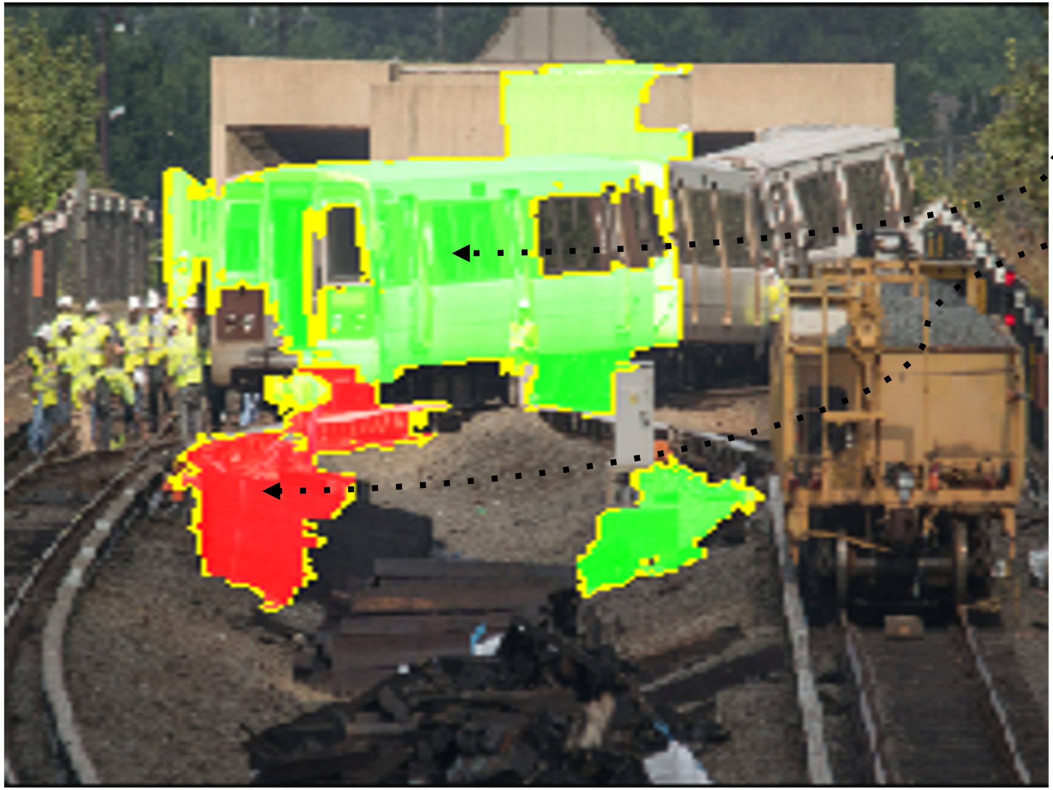
Freddy Lécué, Tanguy Pommellet: Feeding Machine Learning with Knowledge Graphs for Explainable Object Detection. ISWC Satellites 2019: 277-280

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

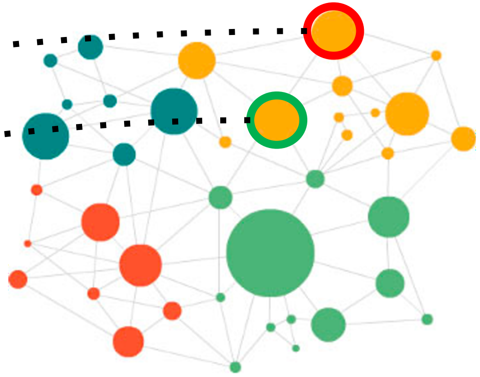
Jiaoyan Chen, Freddy Lécué, Jeff Z. Pan, Ian Horrocks, Huajun Chen: Knowledge-Based Transfer Learning Explanation. KR 2018: 349-358

# Let’s go even Beyond

# Knowledge Graph in Machine Learning (1)

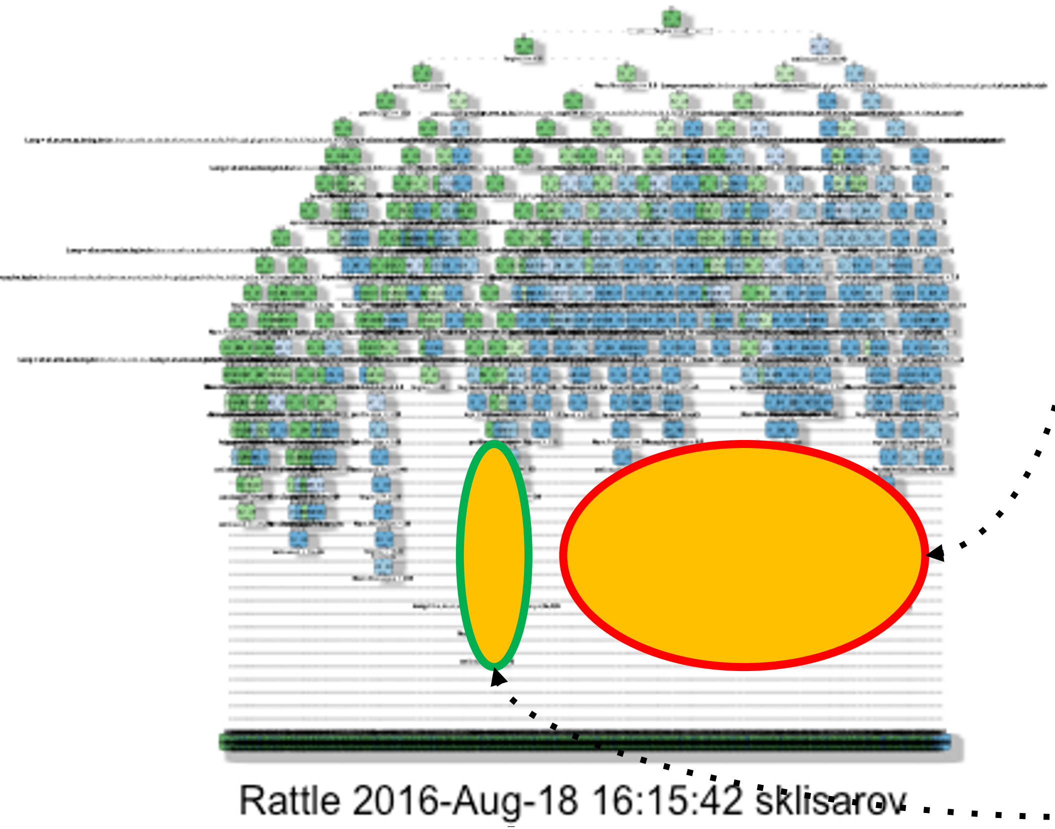


https://stats.stackexchange.com/questions/230581/decision-tree-too-large-to-interpret



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

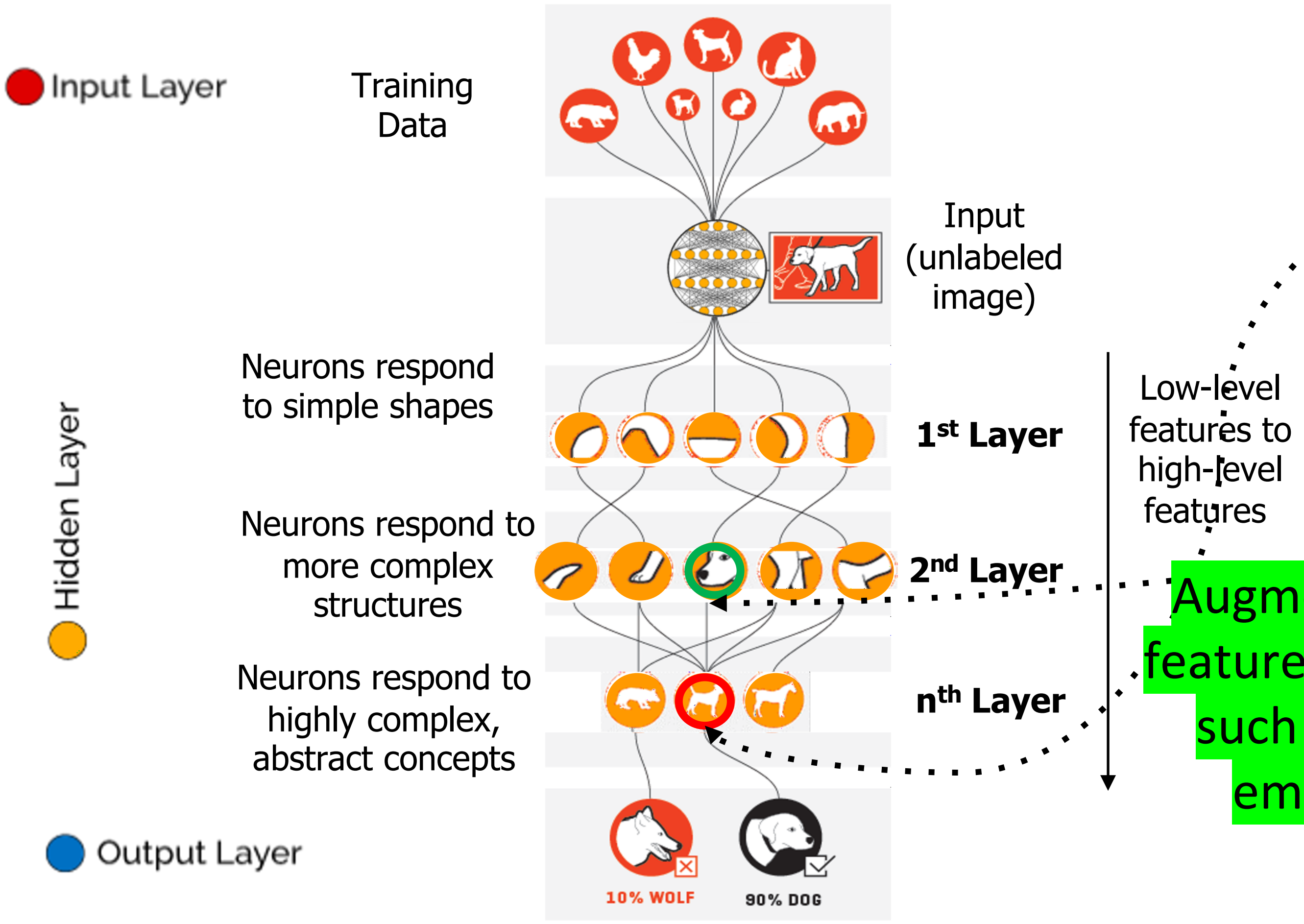
# Knowledge Graph in Machine Learning (2)

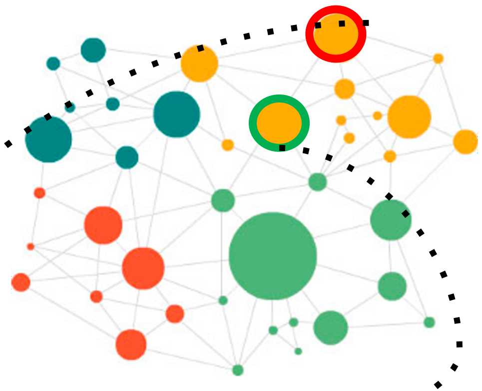



https://stats.stackexchange.com/questions/230581/decision-tree-too-large-to-interpret

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

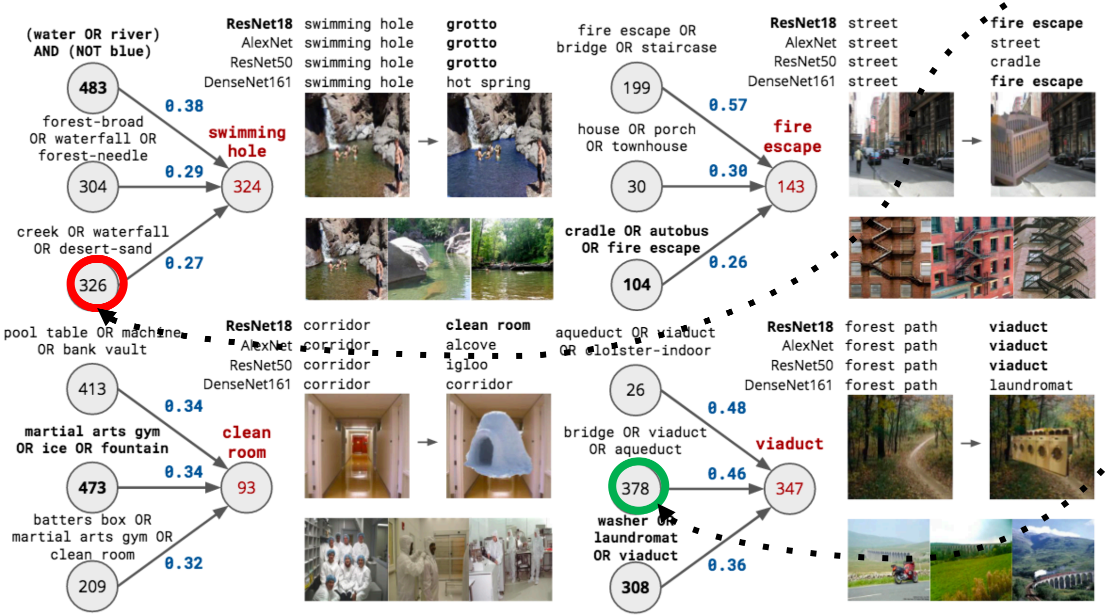
# Knowledge Graph in Machine Learning (3)



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

# Knowledge Graph in Machine Learning (4)

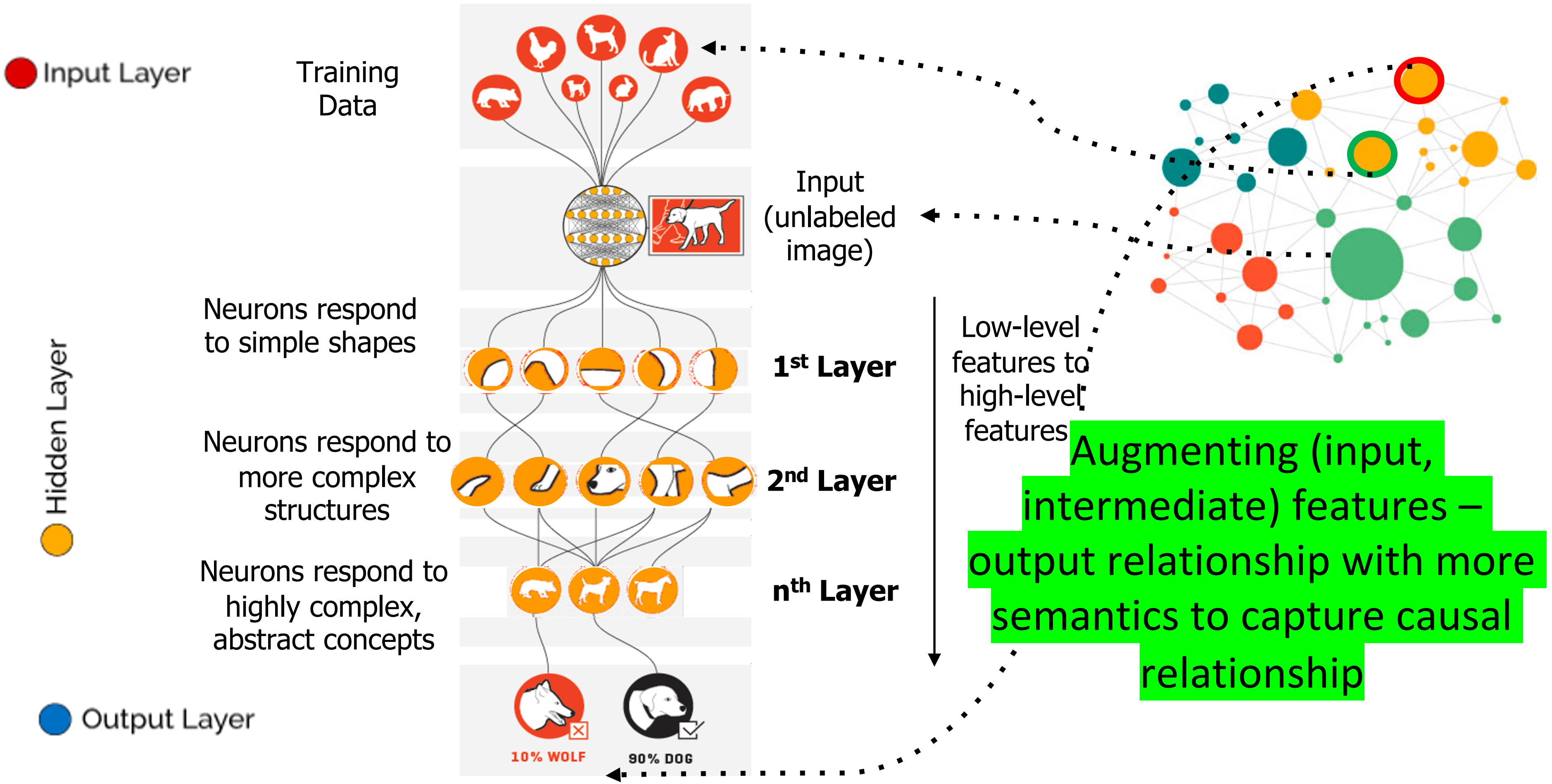



Jesse Mu, Jacob Andreas:Compositional Explanations of Neurons. NeurIPS 2020

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)



# Knowledge Graph in Machine Learning (6)

Description 1: This is an orange train accident

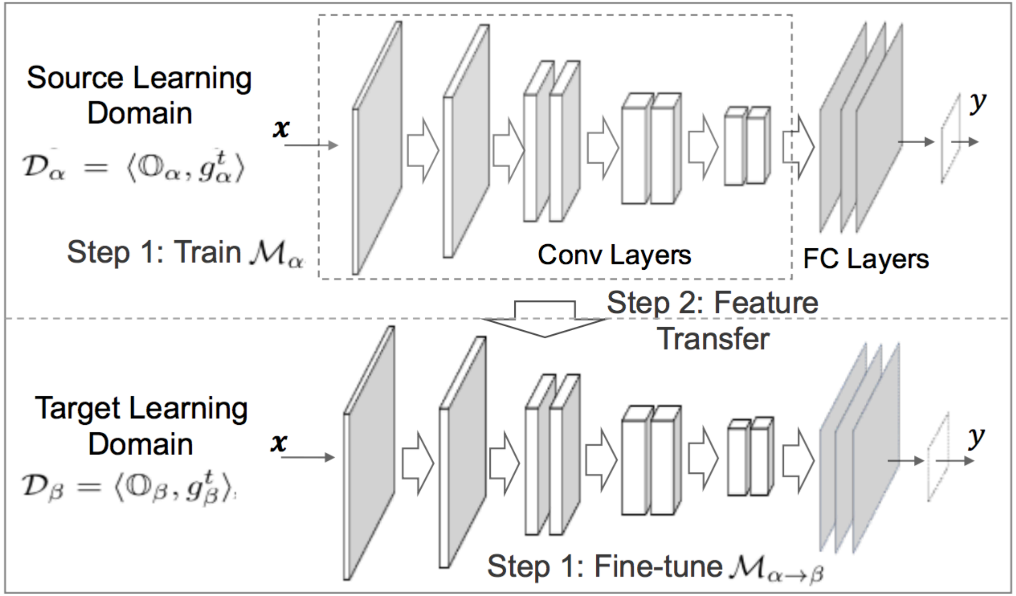
Description 2: This is a train accident between two speed merchant trains of characteristics X43-B and Y33-C in a dry environment

Augmenting models with semantics to support personalized explanation

Description 3: This is a public transportation accident

# 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

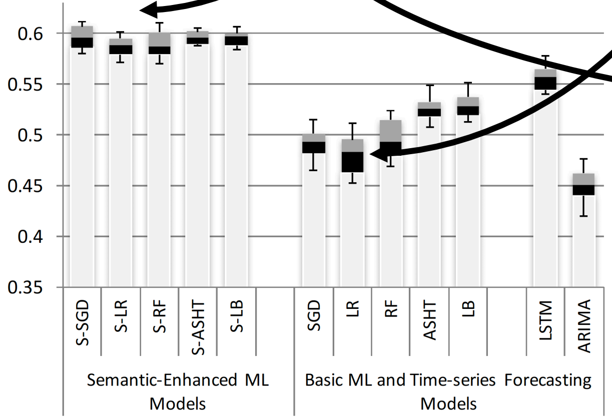



Figure 6: [Beijing Context] Baseline Comparison of Forecasting Macro-F1 Score (Evaluation of Algorithm 1-3), where $\Delta = 6$.

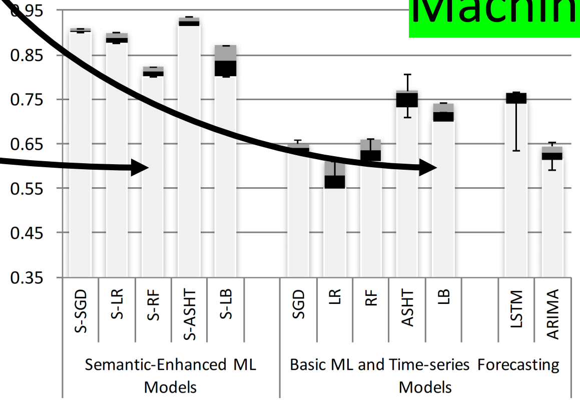



Figure 7: [Dublin Context] Baseline Comparison of Forecasting Macro-F1 Score (Evaluation of Algorithm 1-3), where $\Delta = 6$.

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

# Knowledge Graph in Machine Learning (9)

- Towards more semantic interpretation

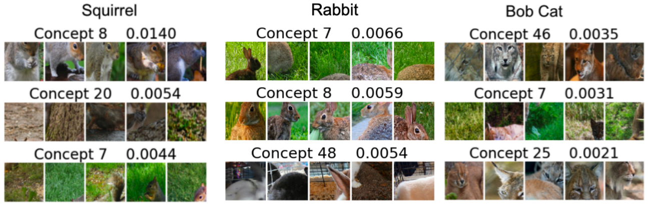
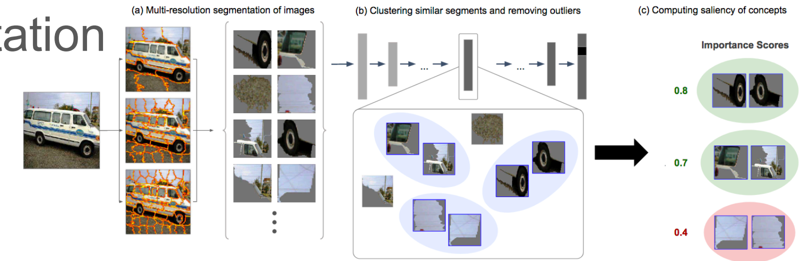


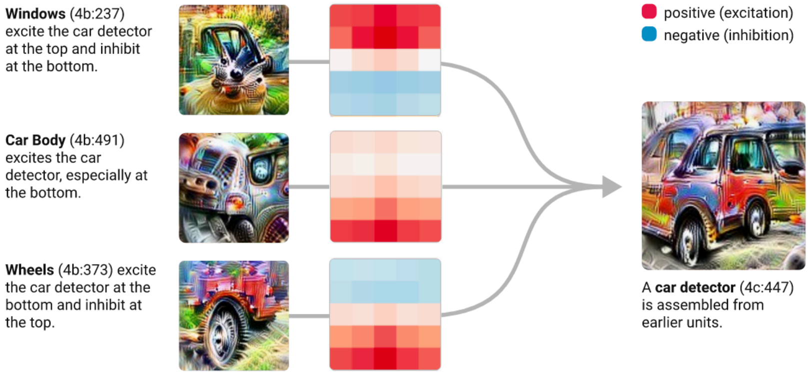

Figure 3: Concept examples with the samples that are the nearest to concept vectors in the activation space in AwA. The per-class ConceptSHAP score is listed above the images.

**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/

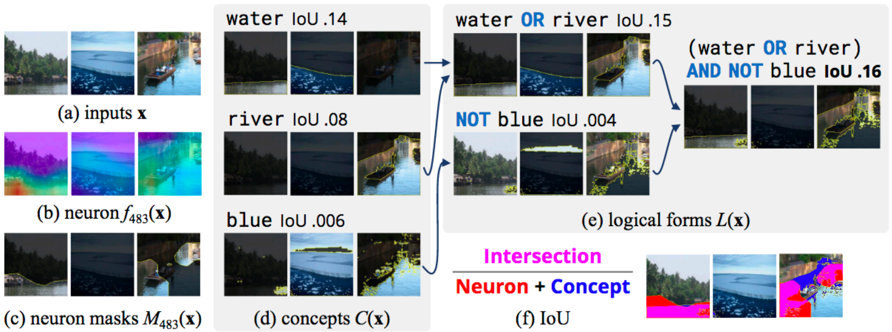



Figure 1: Given a set of inputs (a) and scalar neuron activations (b) converted into binary masks (c), we generate an explanation via beam search, starting with an inventory of primitive concepts (d), then incrementally building up more complex logical forms (e). We attempt to maximize the IoU score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

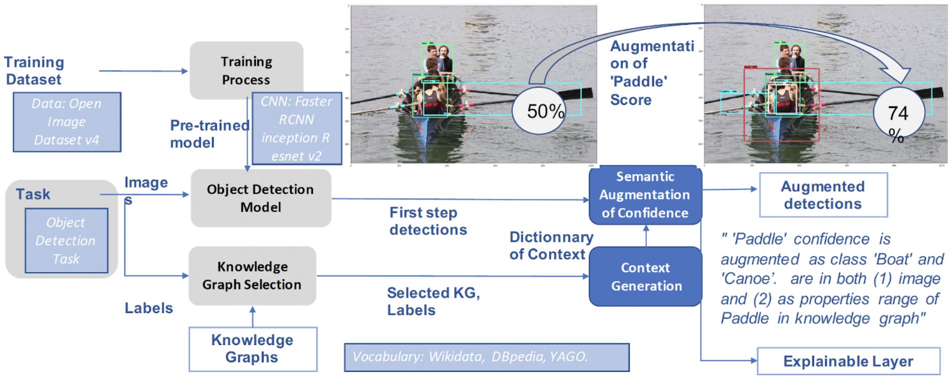
**Compositional Explanations**

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

# Part IV

## XAI Applications and Lessons Learnt

# Explainable Boosted Object Detection – Industry Agnostic



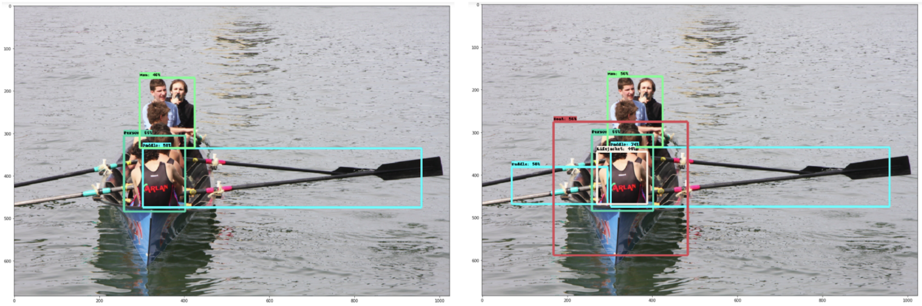



**Fig. 2.** Left image: results from baseline Faster RCNN: Paddle: 50% confidence, Person: 66%, Man: 46%. Right image: results from the semantic augmentation: **Paddle: 74%** confidence, Person: 66%, **Man: 56%**, **Boat: 58% with explanation: Person, Paddle, Water as part of the context in the image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

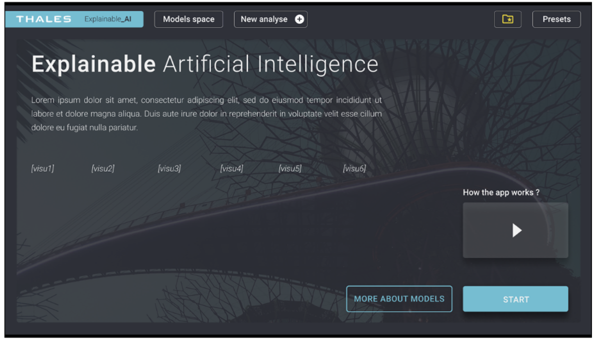
**Challenge:** Object detection is usually performed from a large portfolio of Artificial Neural Networks (ANNs) architectures trained on large amount of labelled data. Explaining object detections is rather difficult due to the high complexity of the most accurate ANNs.

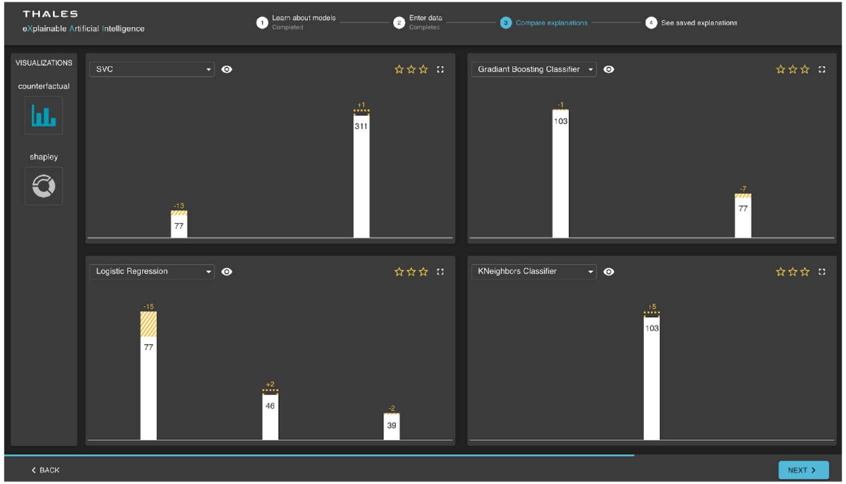
**AI Technology**: Integration of AI related technologies i.e., Machine Learning (Deep Learning / CNNs), and knowledge graphs / linked open data.

**XAI Technology**: Knowledge graphs and Artificial Neural Networks

THALES

# Thales XAI Platform – Industry Agnostic




## Context

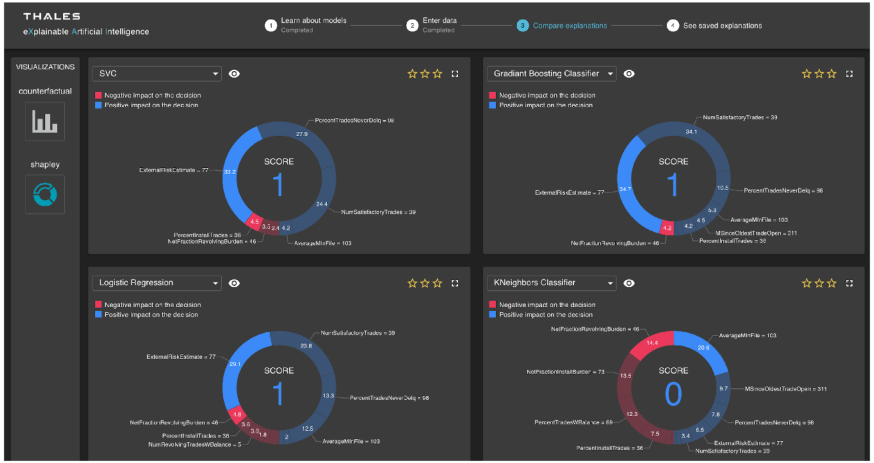
- Explanation in Machine Learning systems has been identified to be the one asset to have for large scale deployment of Artificial Intelligence (AI) in critical systems
- Explanations could be example-based (who is similar), features-based (what is driving decision), or even counterfactual (what-if scenario) 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

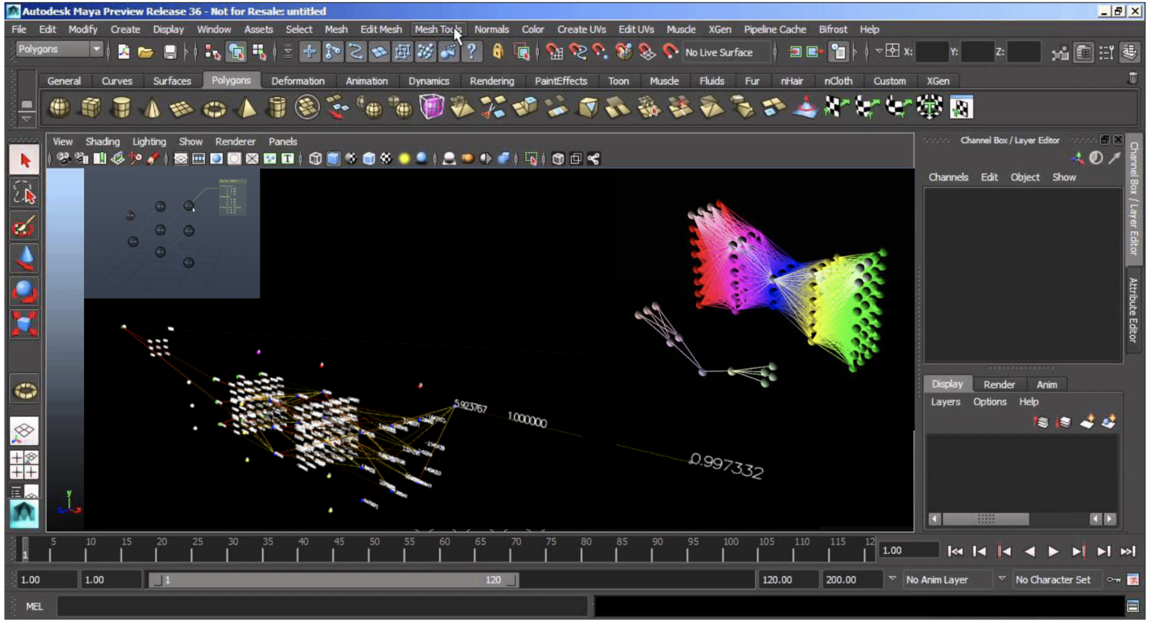
- [AI:ML] Grad-Cam, Shapley, Counter-factual, Knowledge graph




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.

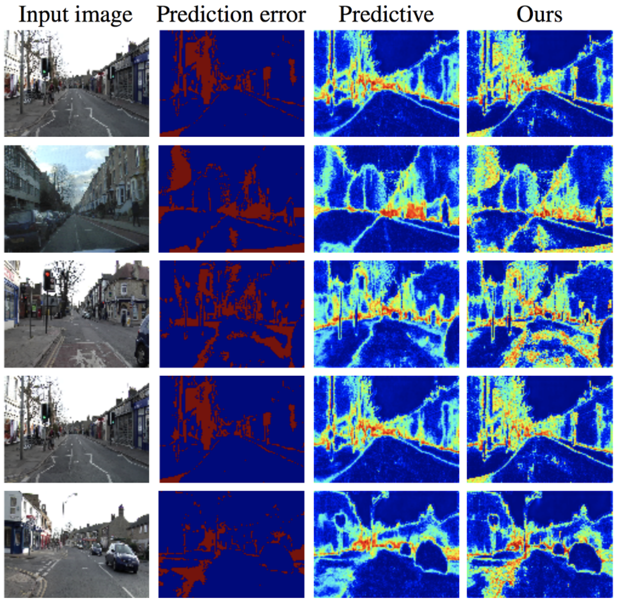
**AI Technology**: Artificial Neural Network

**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




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

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

**XAI Technology**: Deep learning and Epistemic uncertainty

# Explaining Flight Performance – Transportation

**Challenge:** Predicting and explaining aircraft engine performance

**AI Technology**: Artificial Neural Networks

**XAI Technology**: Shapely Values

THALES

# Explainable On-Time Performance – Transportation

KLM / Transavia Flight Delay Prediction

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

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).

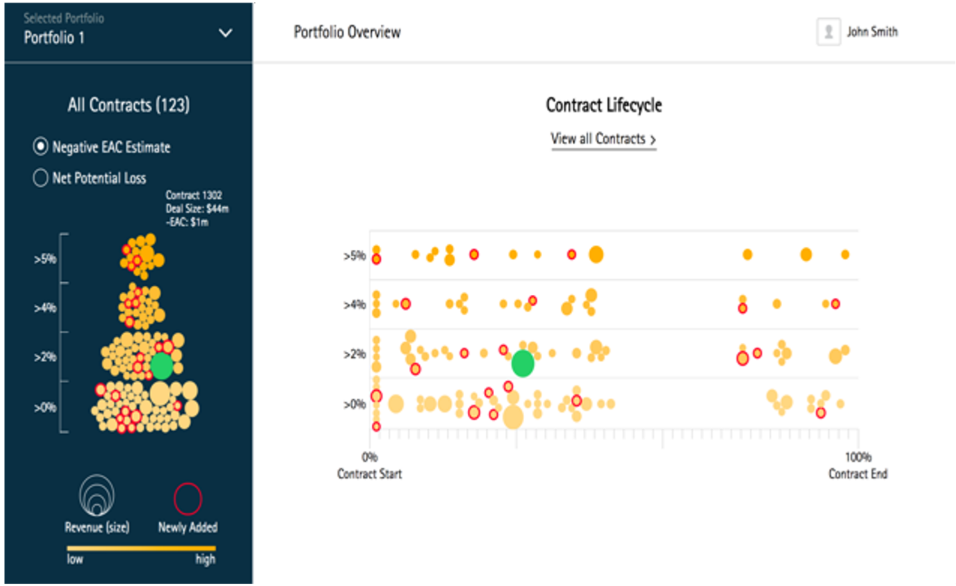
**AI Technology**: Integration of AI related technologies i.e., Machine Learning (Deep Learning / Recurrent neural Network), Reasoning (through semantics-augmented case-based reasoning) and Natural Language Processing for building a robust model which can (1) predict flight delays in minutes, (2) explain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

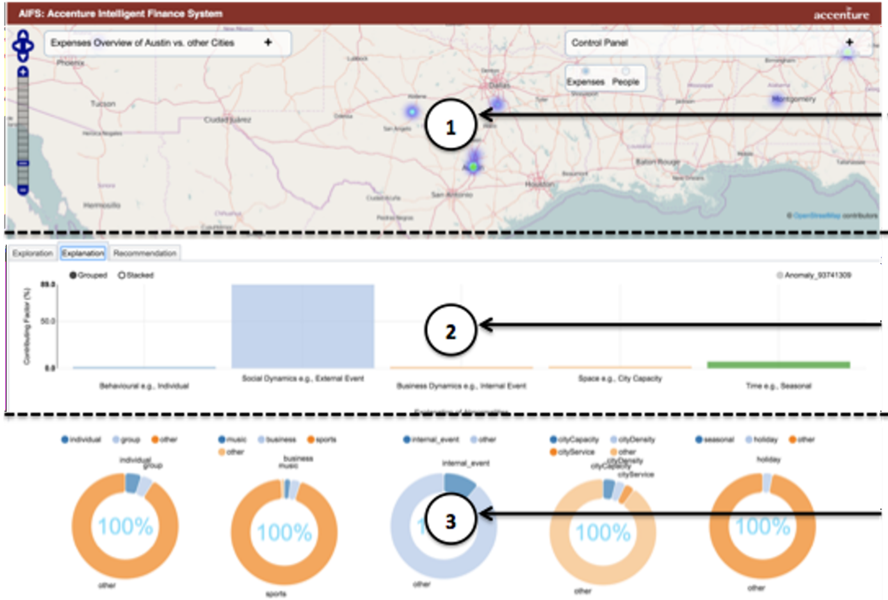
INNOVATION ARCHITECTURE: ACCENTURE LABS

**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)

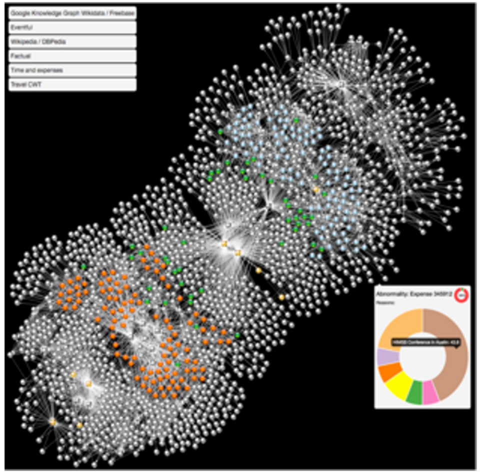



INNOVATION ARCHITECTURE: ACCENTURE LABS

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)



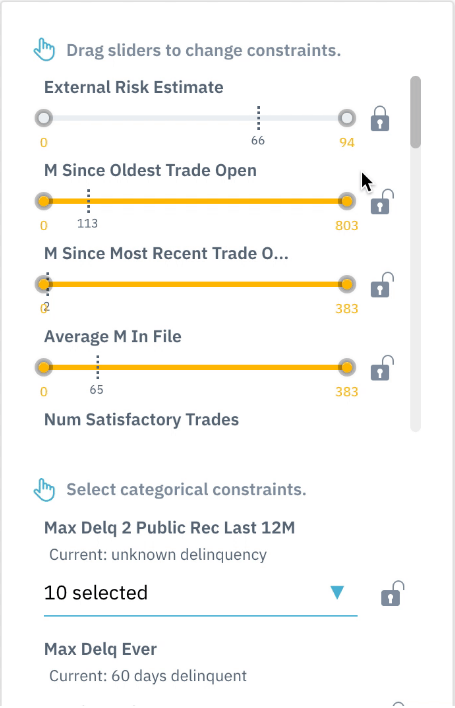
Freddy Lécué, Jiewen Wu: Explaining and predicting abnormal expenses at large scale using knowledge graph based reasoning. J. Web Sem. 44: 89-103 (2017)

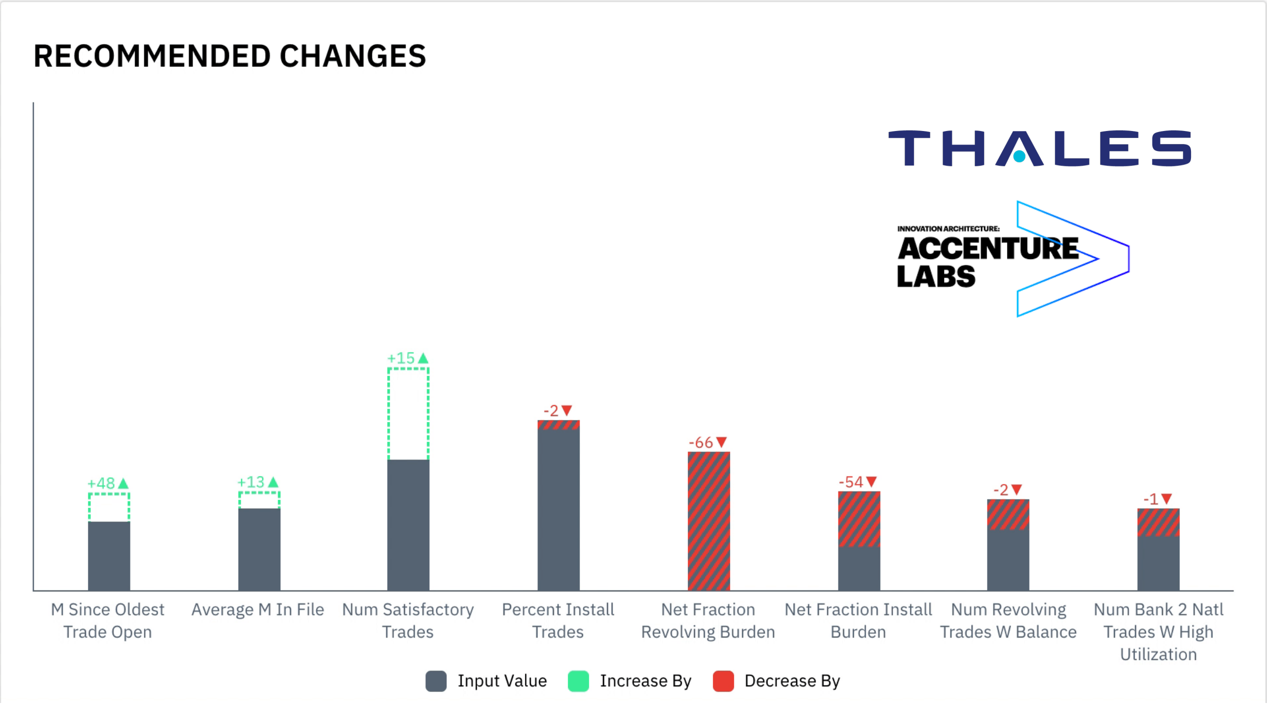
**Challenge:** Predicting and explaining abnormally employee expenses (as high accommodation price in 1000+ cities).

**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

# 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/

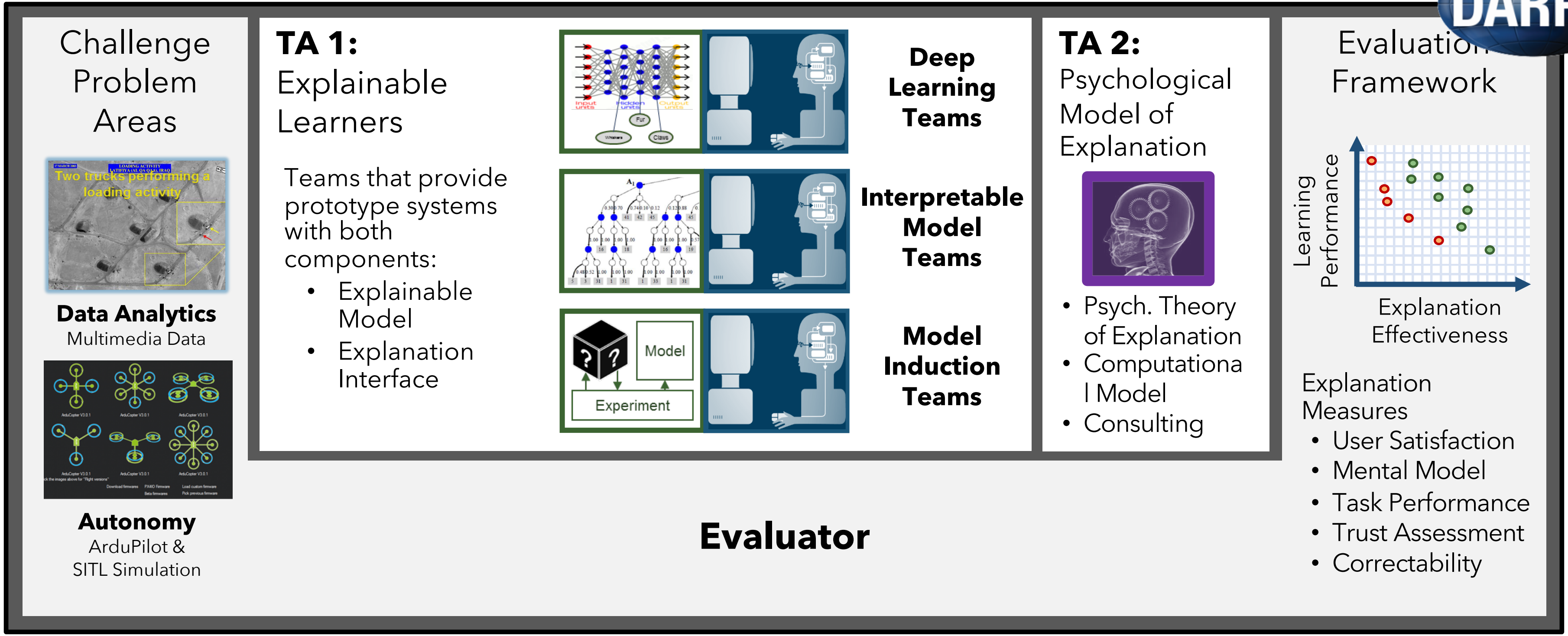
## Challenge:

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

| System Robustness | Certificability | Explicability & Interpretability | Privacy by design |
|---|---|---|---|
| - To biased data<br>- Of algorithm<br>- To change<br>- To attacks | - Structural warranties<br>- Risk auto evaluation<br>- External audit | | - Differential privacy<br>- Homomorphic coding<br>- Collaborative learning<br>- 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 **<- Role of Semantics**
- 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 (@freddylecue)**
- Slides: **https://tinyurl.com/9ahdbtm4**
- Extended version (youtube link): https://www.youtube.com/watch?v=uFF1Ul1oM88
- To try Thales XAI Platform , please send an email to **freddy.lecue@thalesgroup.com**

THALES

Inria
INVENTEURS DU MONDE NUMÉRIQUE