



北京航空航天大學
BEIHANG UNIVERSITY

可解释人工智能

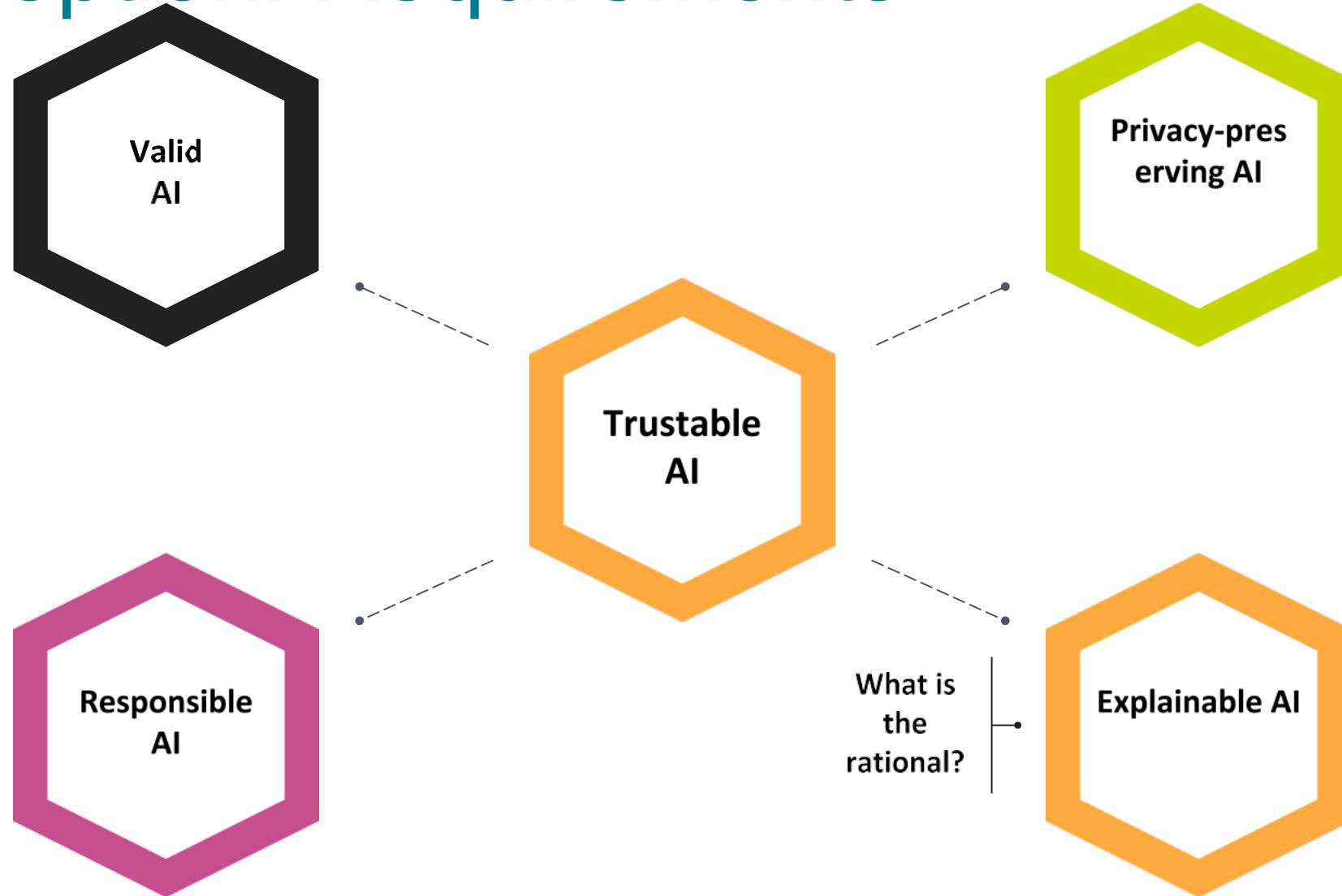
人工智能研究院

主讲 沙磊

Agenda

- Part I: Introduction and Motivation
 - Motivation, Definitions, Properties, Evaluation
 - Challenges for Explainable AI @ Scale
- Part II: Explanation in AI (not only Machine Learning!)
 - From Machine Learning to Knowledge Representation and Reasoning and Beyond
- Part III: Explainable Machine Learning (from a Machine Learning Perspective)
- Part IV: Explainable Machine Learning (from a Knowledge Graph Perspective)

AI Adoption: Requirements





Explanation - From a Business Perspective

Business to Customer AI



Gary Chavez added a photo you might be in.
about a minute ago · 



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



DYLAN FUGETT

Prior Offense

1 attempted burglary

Subsequent Offenses

3 drug possessions

LOW RISK

3

BERNARD PARKER

Prior Offense

1 resisting arrest without violence

Subsequent Offenses

None

HIGH RISK

10

Fugett was rated low risk after being arrested with cocaine and marijuana. He was arrested three times on drug charges after that.

... but not only Critical Systems (2)

Finance:

- Credit scoring, loan approval
- Insurance quotes



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



Oliver Ralph MAY 16, 2017

□ 24

... 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.

[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
y lou@linkedin.com

Johannes Gehrke
Microsoft
johannes@microsoft.com

Paul Koch
Microsoft Research
paulkoch@microsoft.com

Marc Sturm
New York-Presbyterian Hospital
mas9161@nyp.org

Noémie Elhadad
Columbia University
noemie.elhadad@columbia.edu

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

Black-box AI creates business risk for Industry

Bloomberg Businessweek

Apple Card's Gender-Bias Claims Look Familiar to Old-School Banks

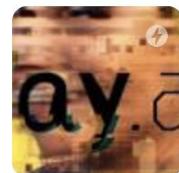
Updated on November 12, 2019, 4:23 AM



BBC NEWS

Tay: Microsoft issues apology over racist chatbot fiasco

Sep 22, 2017



MIT News

Study finds gender and skin-type bias in commercial AI systems

Feb 12, 2018



Missouri S&T News and Research

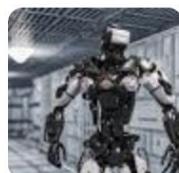
After Uber, Tesla incidents, can artificial intelligence be trusted?

Apr 10, 2018

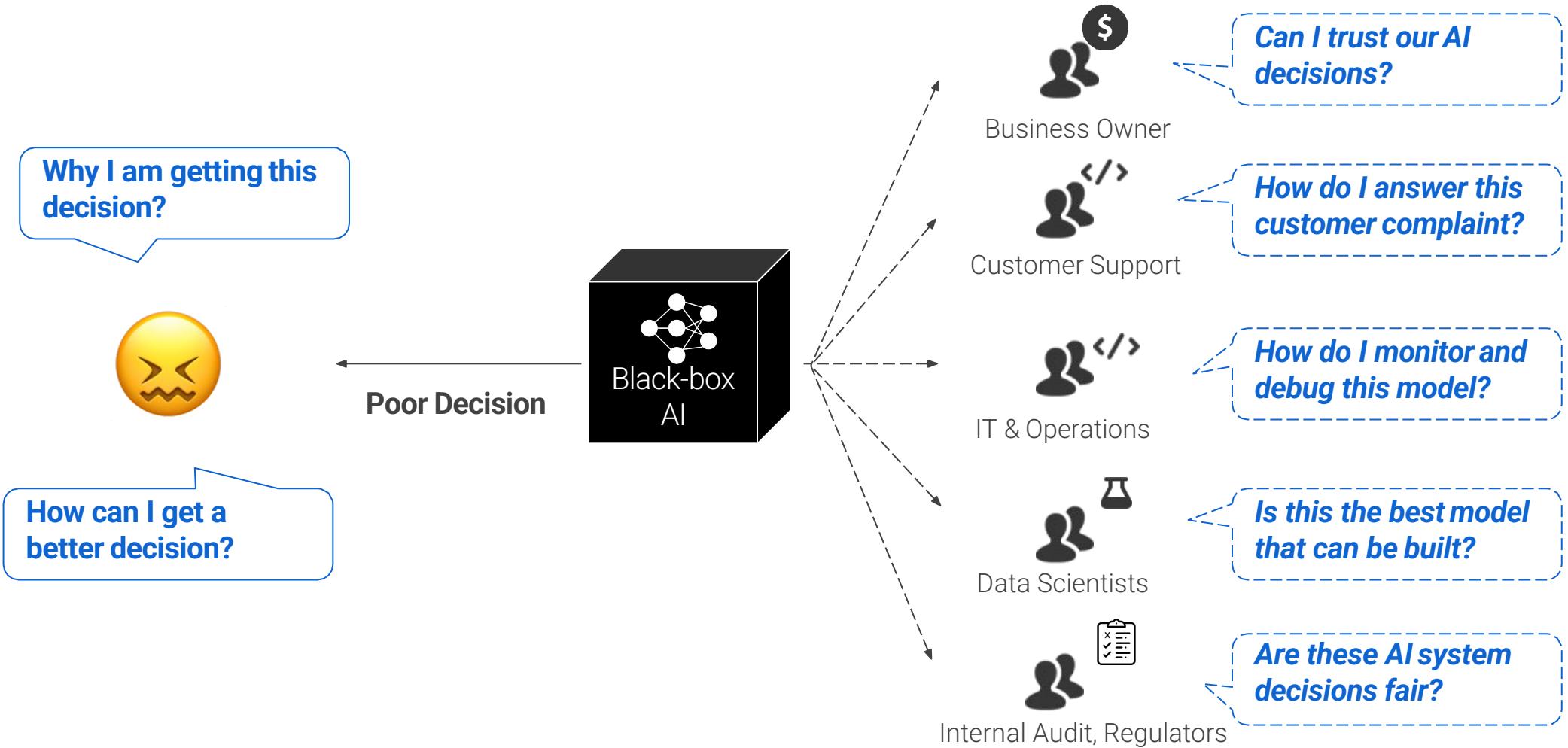


Guilty! AI Is Found to Perpetuate Biases in Jailing

1 day ago



Black-box AI creates confusion and doubt





Explanation - From a Model Perspective

Why Explainability: Debug (Mis-)Predictions

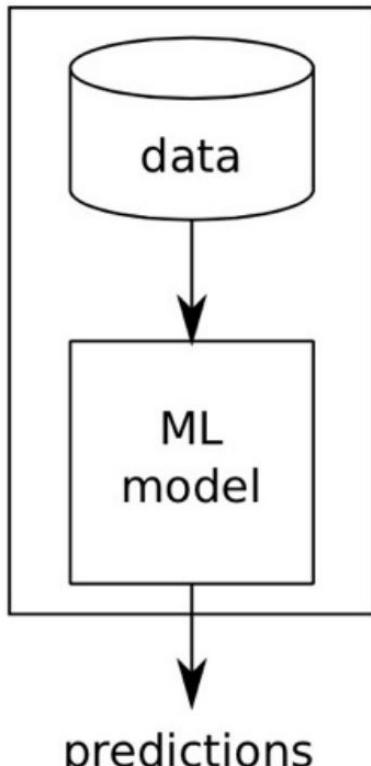


Top label: “**clog**”

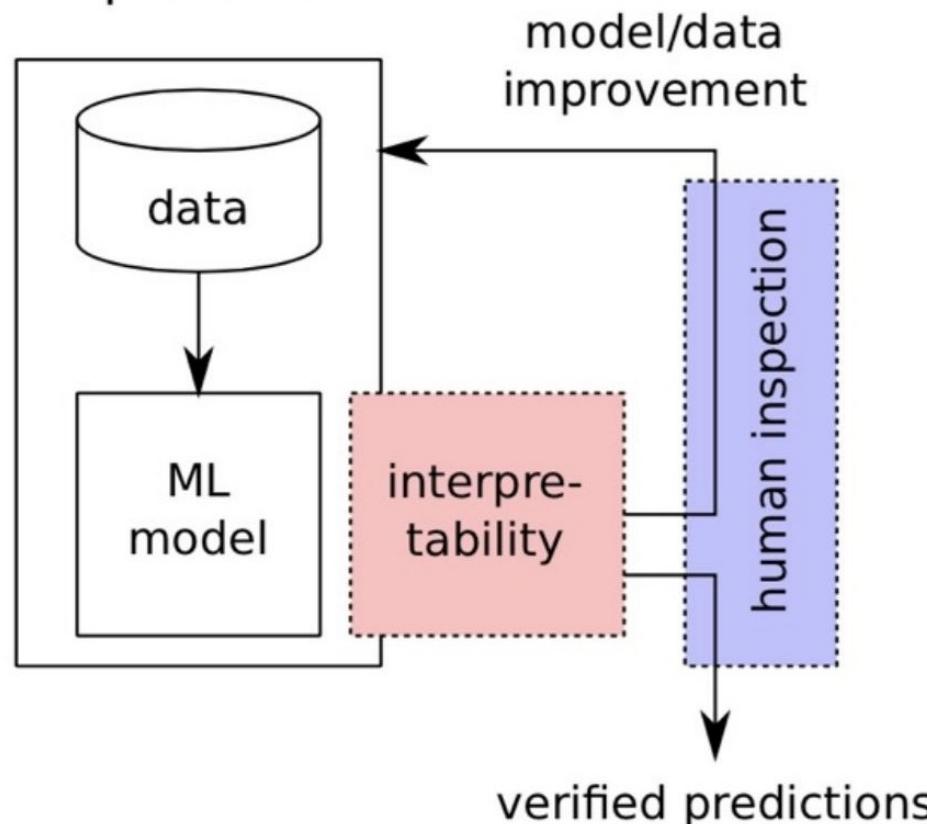
Why did the network label this image as “clog”?

Why Explainability: Improve ML Model

Standard ML



Interpretable ML



Generalization error

Generalization error + human experience

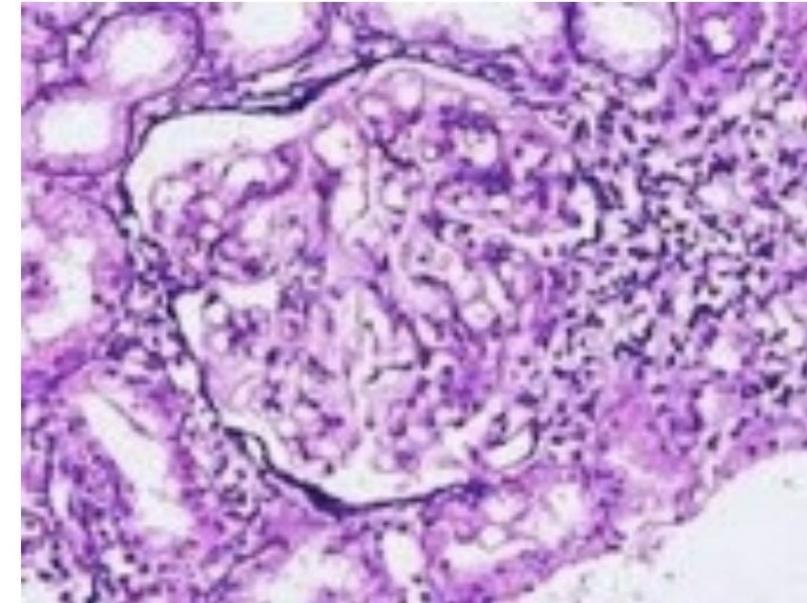
Why Explainability: Verify the ML Model / System

Wrong decisions can be costly
and dangerous

*“Autonomous car crashes,
because it wrongly recognizes ...”*



*“AI medical diagnosis system
misclassifies patient’s disease ...”*

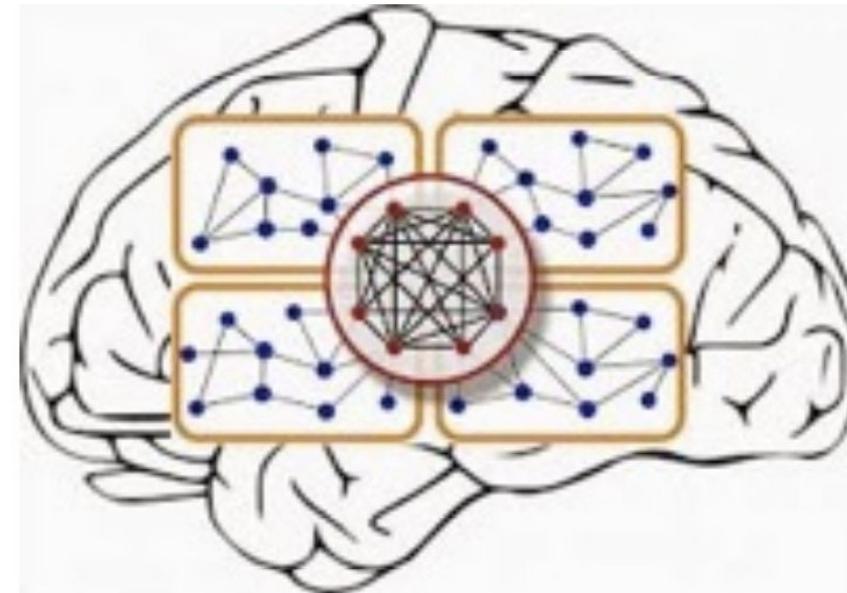


Why Explainability: Learn New Insights

“It's not a human move. I've never seen a human play this move.” (Fan Hui)

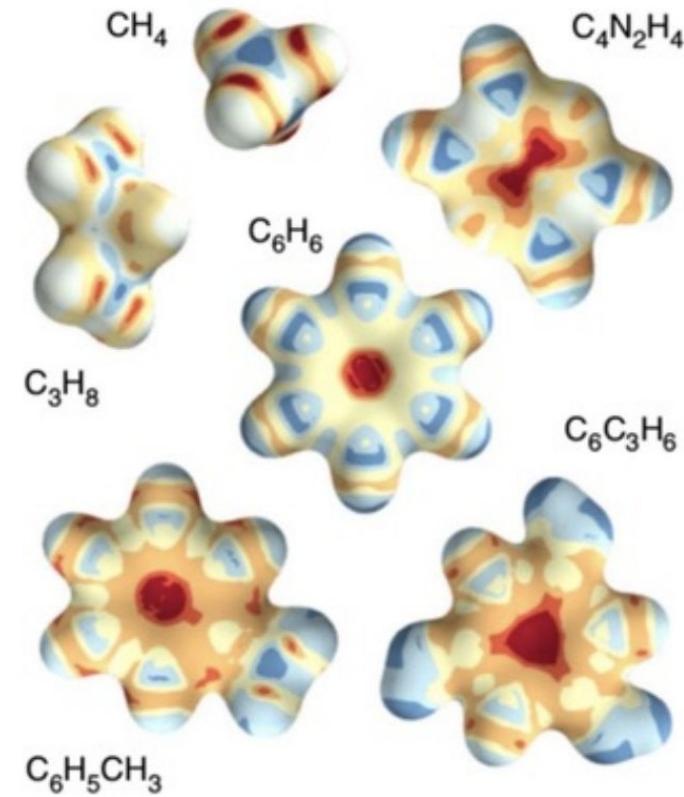
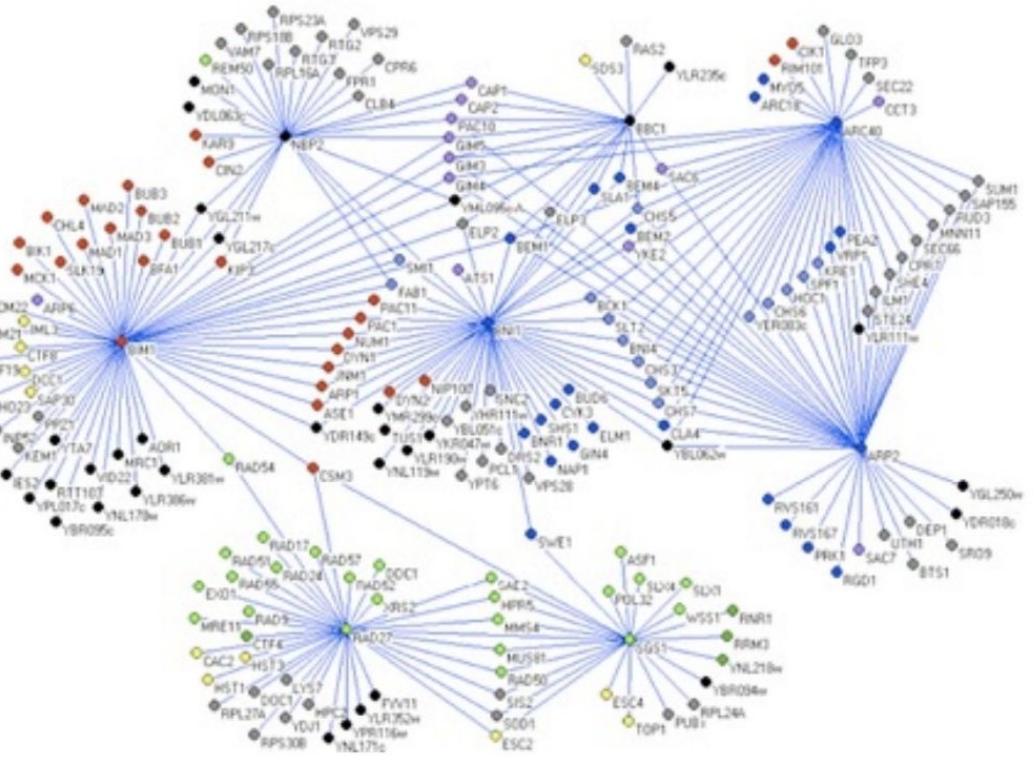


Old promise:
“Learn about the human brain.”



Why Explainability: Learn Insights in the Sciences

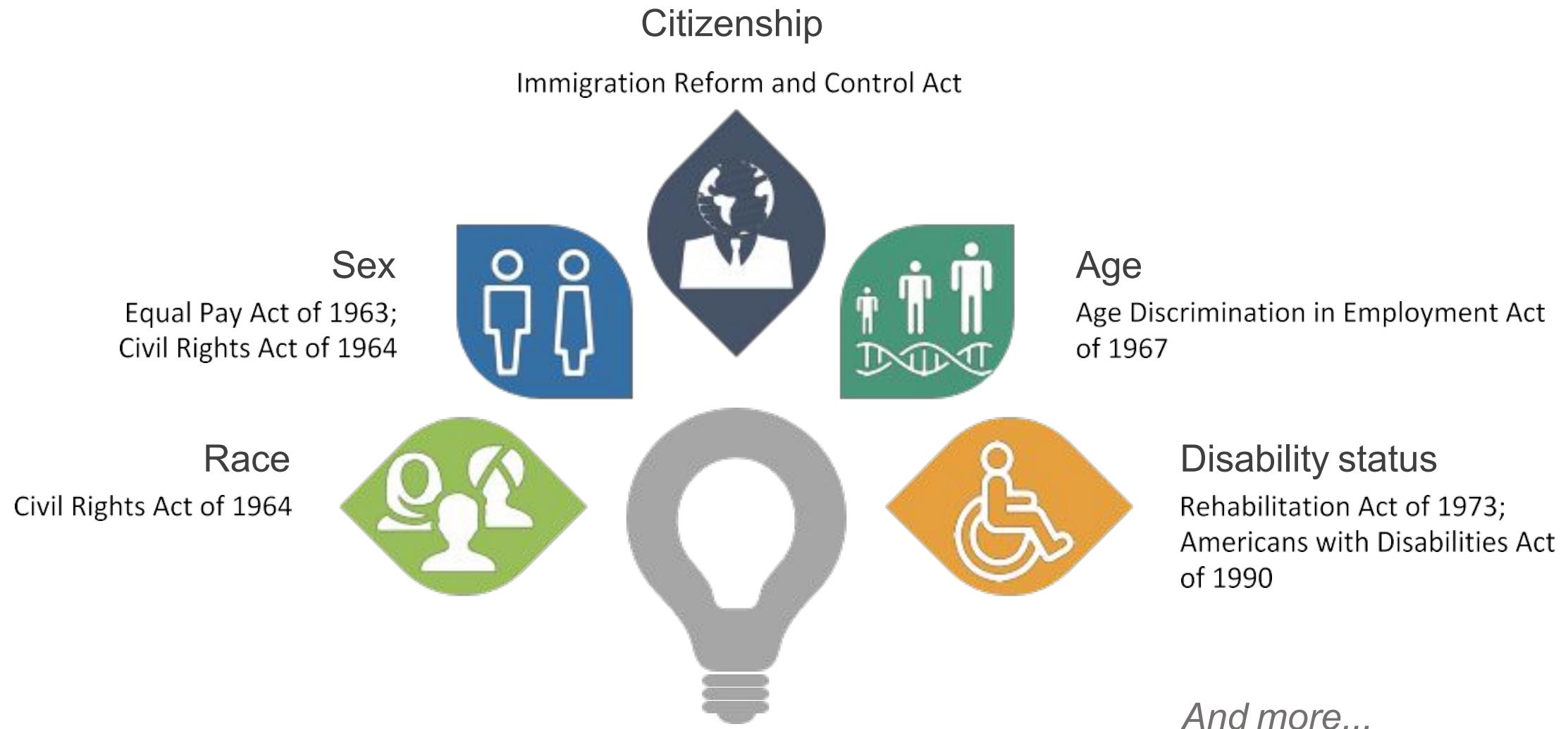
Learn about the physical / biological / chemical mechanisms.
(e.g. find genes linked to cancer, identify binding sites ...)





Explanation - From a Regulatory Perspective

Why Explainability: Laws against Discrimination



Fairness



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

Privacy



Transparency

Explainability

GDPR Concerns Around Lack of Explainability in AI

“

*Companies should commit to ensuring systems that could fall under GDPR, including AI, will be compliant. The threat of **sizeable fines of €20 million or 4% of global turnover** provides a sharp incentive.*

*Article 22 of GDPR empowers individuals with the **right to demand an explanation of how an AI system made a decision that affects them.***

”

- European Commission



Andrus Ansip ✅
@Ansip_EU

You have the right to be informed about an automated decision and ask for a human being to review it, for example if your online credit application is refused.
#EUdataP #GDPR #AI #digitalrights #EUandMe europa.eu/nN77Dd



8:30 AM - 7 Sep 2018

VP, European Commission

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.

Recital 71

Profiling*

Fai

cy

¹ The data subject should have the right not to be subject to a decision, which may include a measure, evaluating personal aspects relating to him or her which is based solely on automated processing and which produces legal effects concerning him or her or similarly significantly affects him or her, such as automatic refusal of an online credit application or e-recruiting practices without any human intervention. ² Such processing includes ‘profiling’ that consists of any form of automated processing of personal data evaluating the personal aspects relating to a natural person, in particular to analyse or predict aspects concerning the data subject’s performance at work, economic situation, health, personal preferences or interests, reliability or behaviour, location or movements, where it produces legal effects concerning him or her or similarly significantly affects him or her. ³ However, decision-making based on such processing,



Transparency

Explainability

Why Explainability: 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 zipcode.

SR 11-7 and OCC regulations for Financial Institutions

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**”

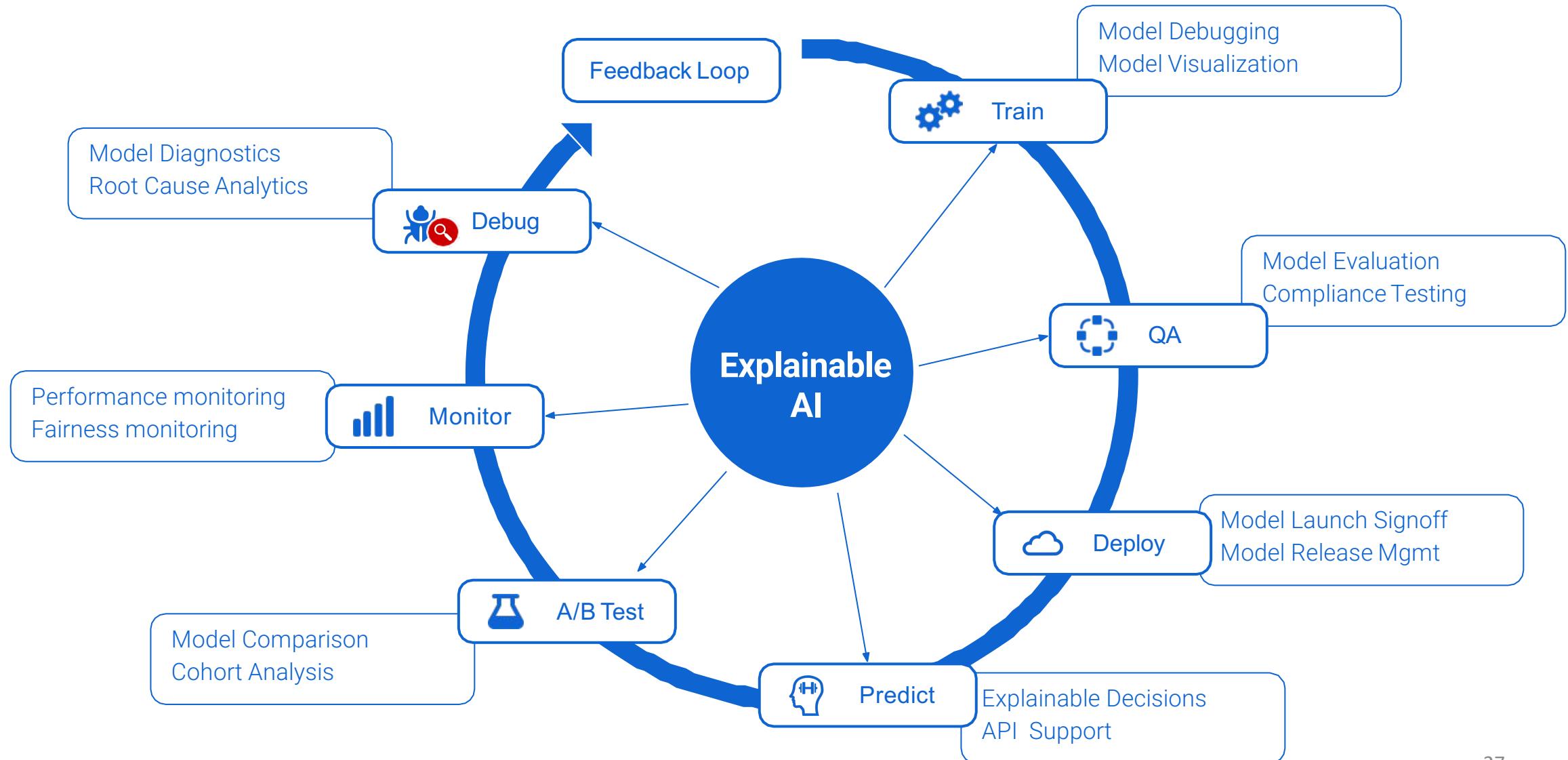
In fact, SR14-03 explicitly calls for all models used for Dodd-Frank Act Company-Run Stress Tests must fall under the purview of Model Risk Management.

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.

JOHN HILL
GLOBAL HEAD OF MODEL RISK GOVERNANCE, **CREDIT SUISSE**

// In the current regulatory environment, model validation policies must be fully compliant with the requirements of SR11-7. While SR11-7 officially applies to US conforming bank and non-US banks doing business in the US, many European financial firms have adopted SR11-7 as their standard as well. **//**

“Explainability by Design” for AI products

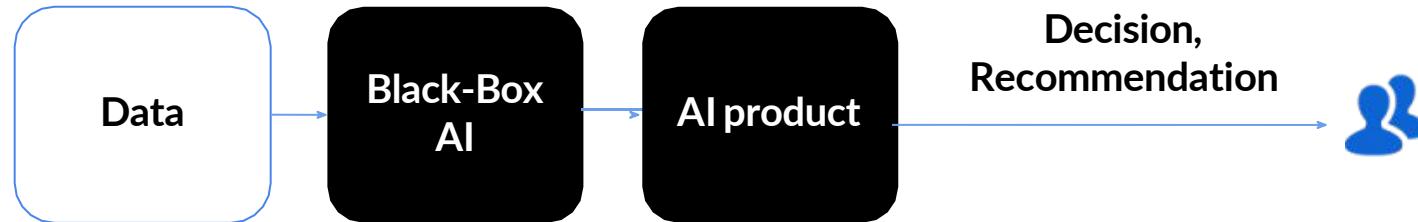




Explanation - In a Nutshell

What is Explainable AI?

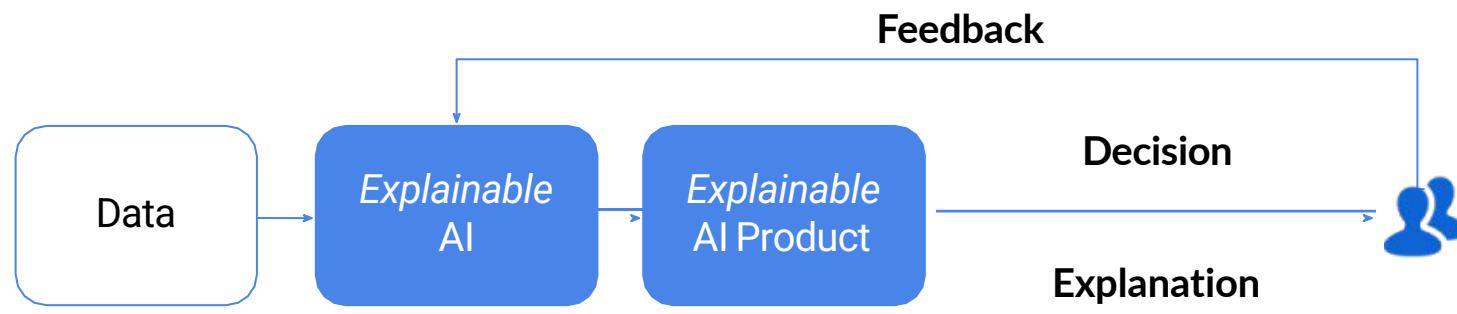
Black Box AI



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

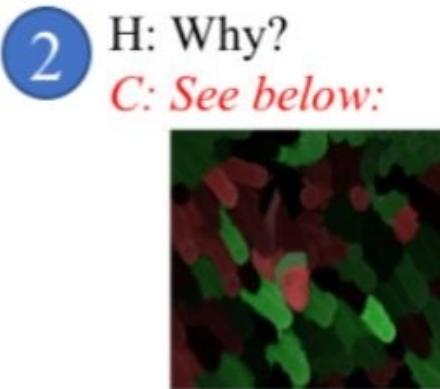
- I understand why
- I understand why not
- I know why you succeed or fail
- I understand, so I trust you

Example of an End-to-End XAI System



ML Classifier

C: I predict FISH



Green regions argue for FISH, while RED pushes towards DOG. There's more green.

3 H: (Hmm. Seems like it might be just recognizing anemone texture!) Which training examples are most influential to the prediction?

C: These ones:



4 H: What happens if the background anemones are removed? E.g.,



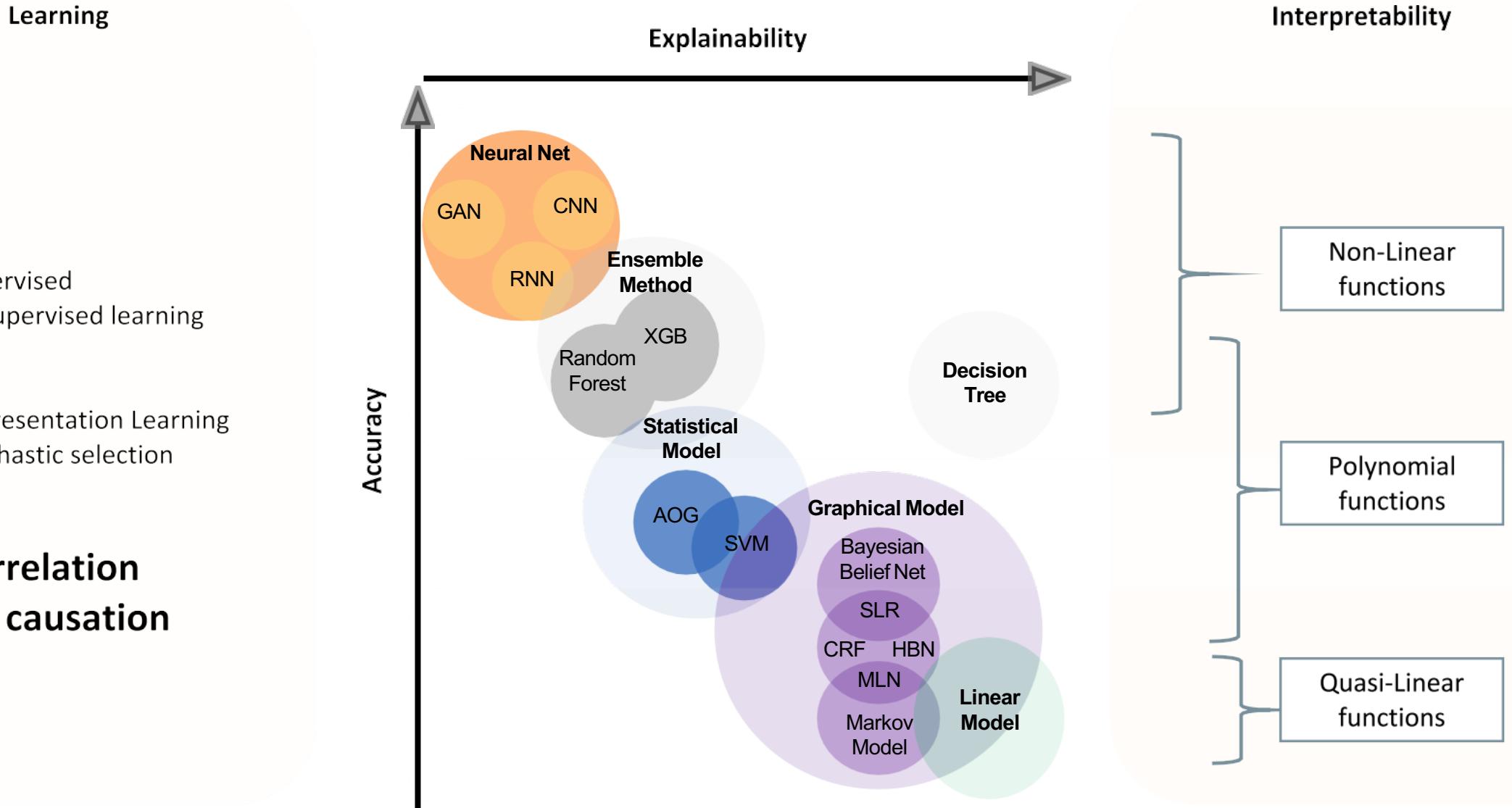
C: I still predict FISH, because of these green superpixels:



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

How to Explain? Accuracy vs. Explainability

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



XAI Definitions - Explanation vs. Interpretation

explanation | ɛksplə'neɪʃ(ə)n |

noun

Oxford Dictionary of English

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

interpret | ɪn'tə:prɪt |

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

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

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

2000 rows
all told

Field
names

One row
(4 fields)

Tabular

Images



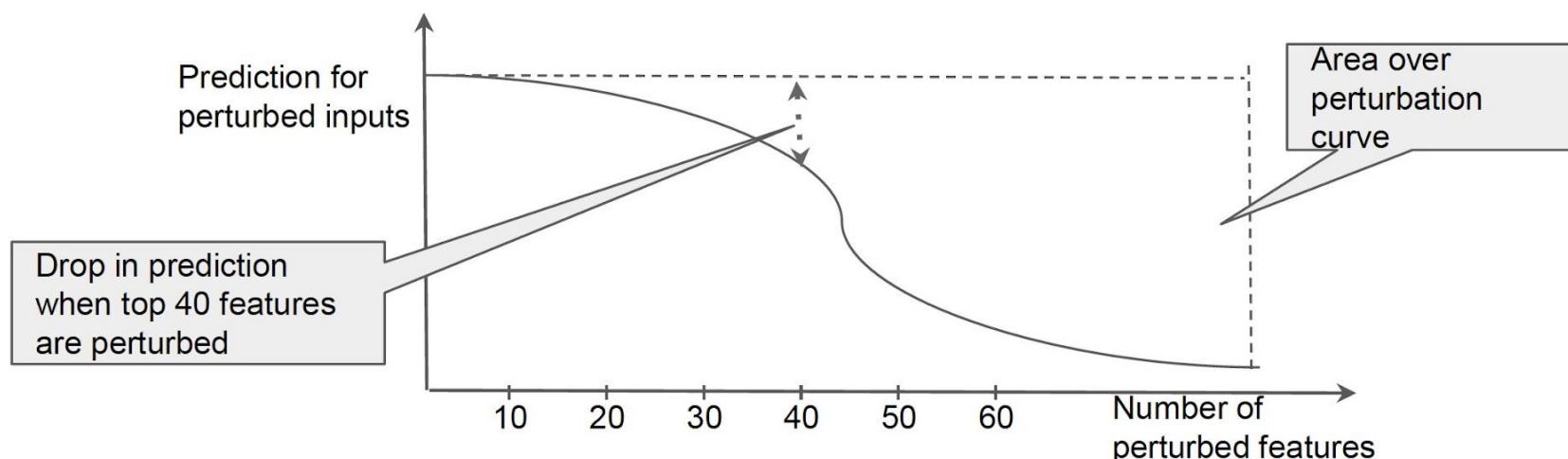
Text

© Bernard Castlein / naturepl.com

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

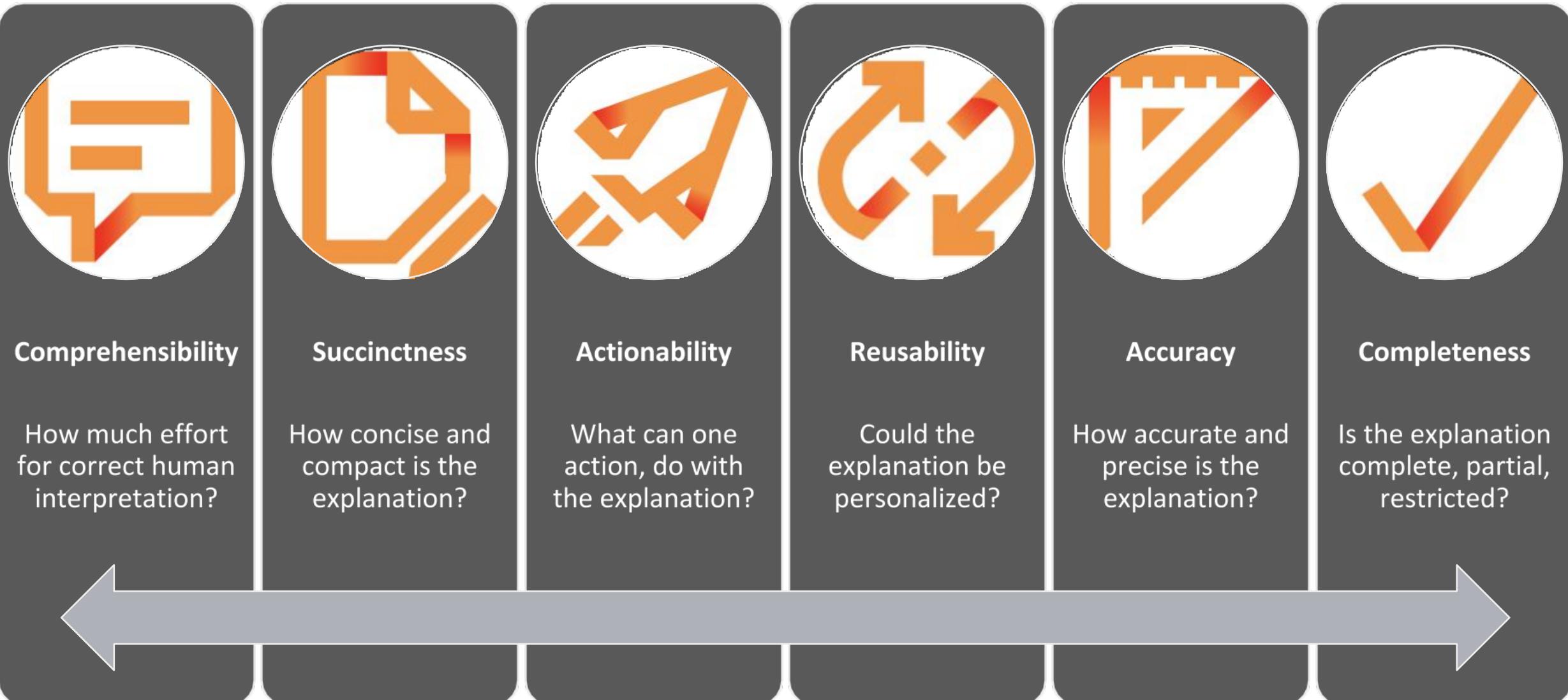
Evaluation criteria for Explanations [Miller, 2017]

- Truth & probability
- Usefulness, relevance
- Coherence with prior belief
- Generalization

Cognitive chunks = basic explanation units (for different explanation needs)

- Which basic units for explanations?
- How many?
- How to compose them?
- Uncertainty & end users?

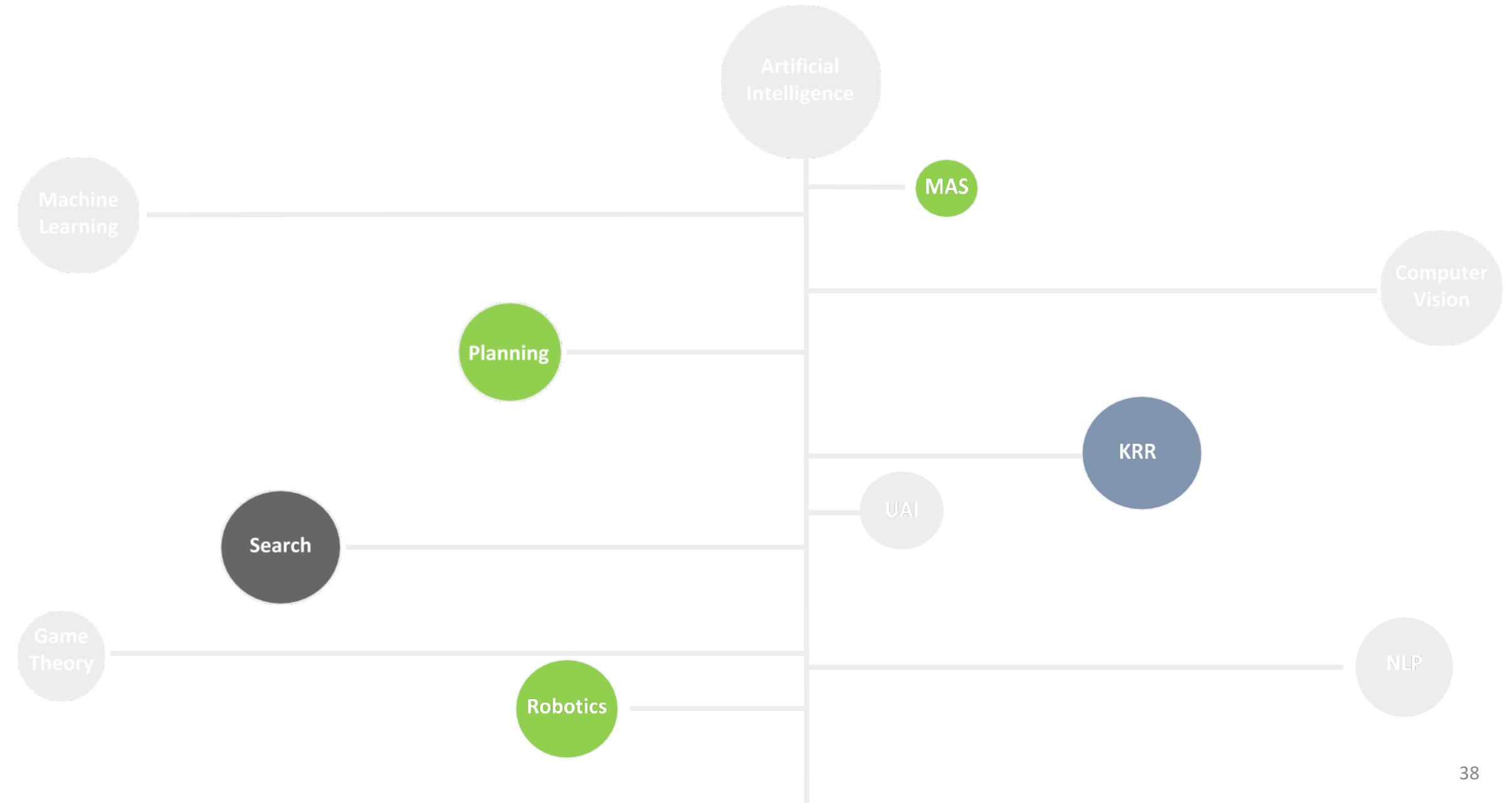
Evaluation (3) - XAI: One Objective, Many Metrics



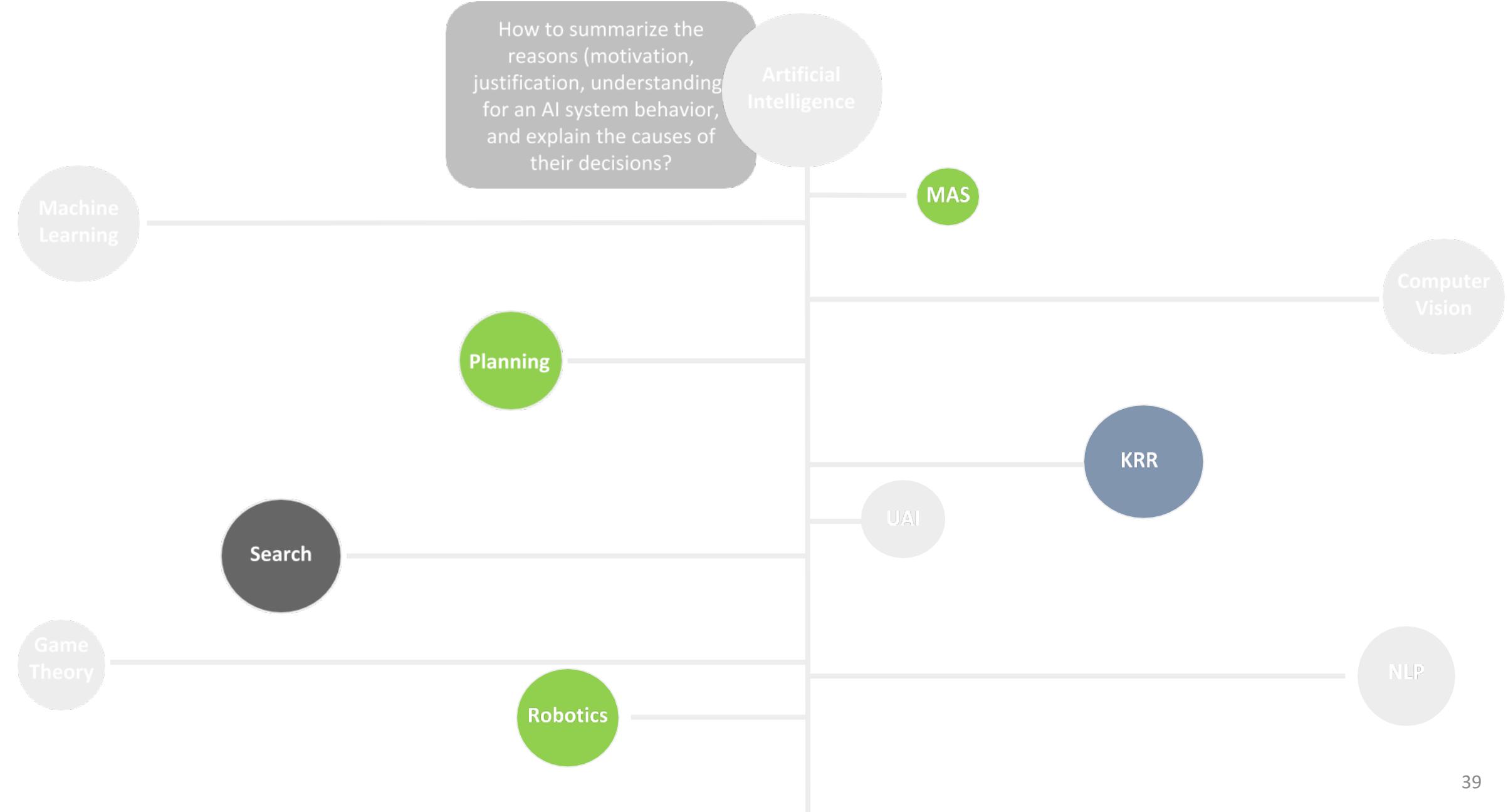


Explanation in AI
(not only Machine
Learning!)

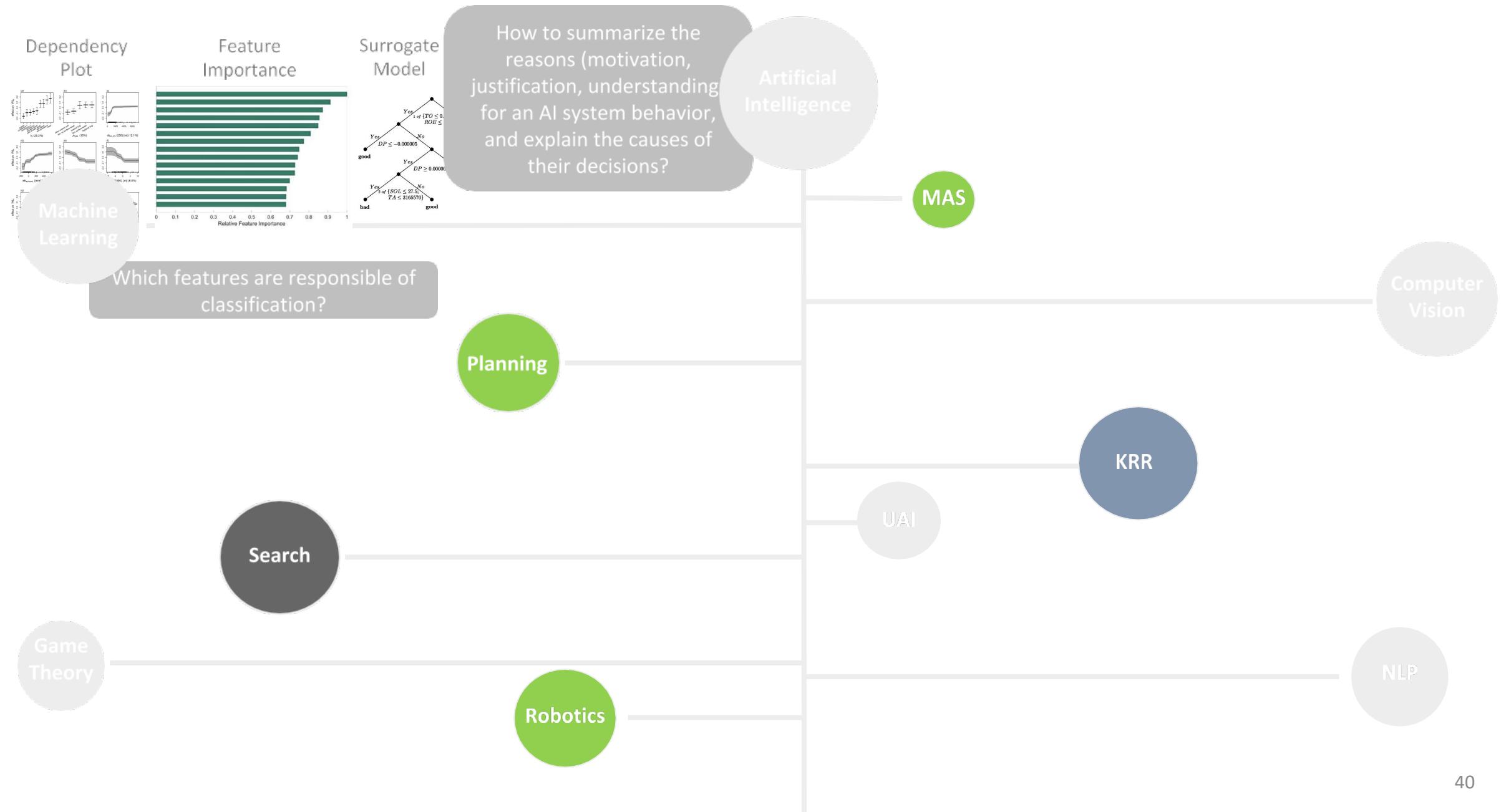
XAI: One Objective, Many ‘AI’s, Many Definitions, Many Approaches



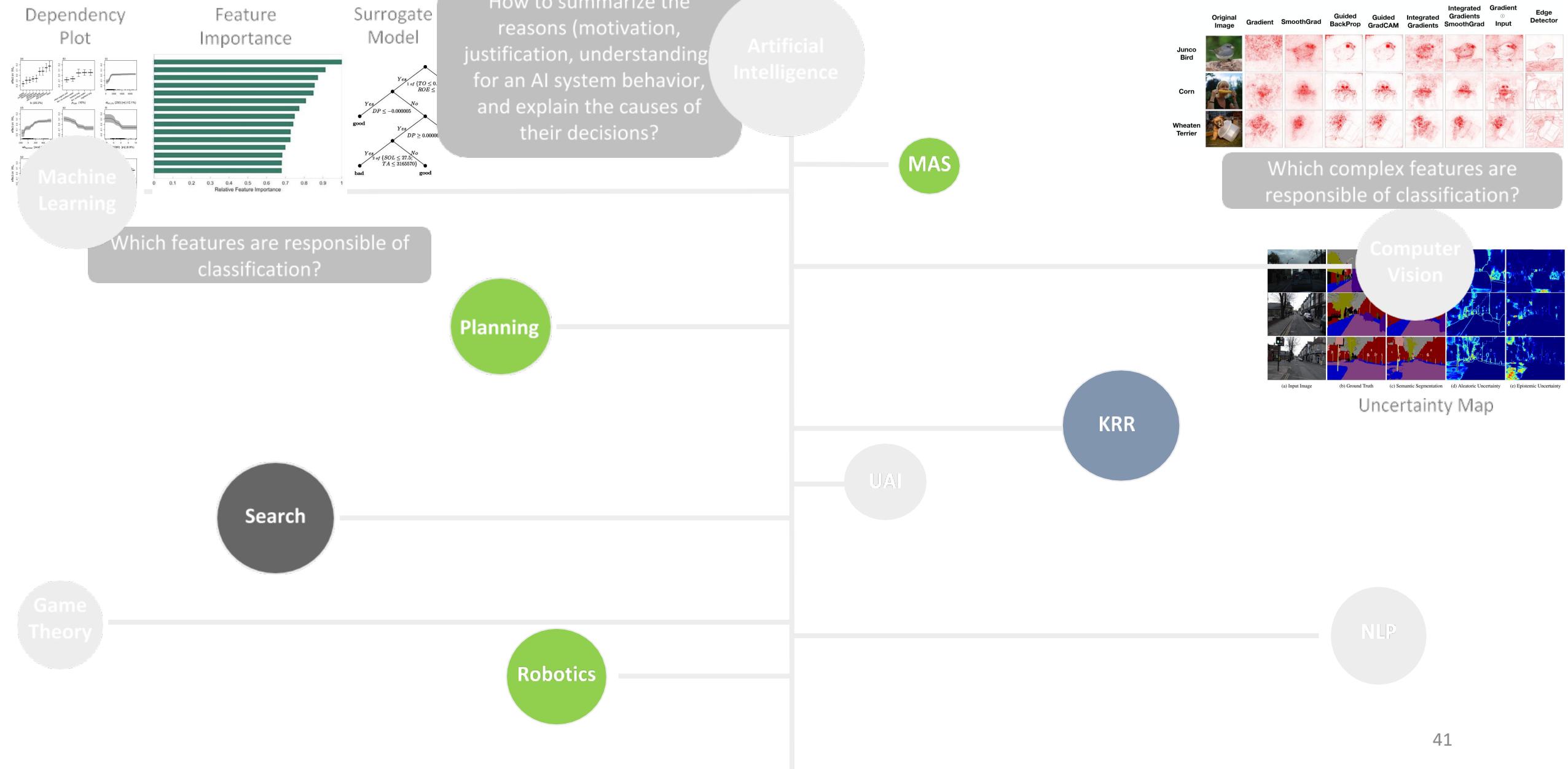
XAI: One Objective, Many ‘AI’s, Many Definitions, Many Approaches



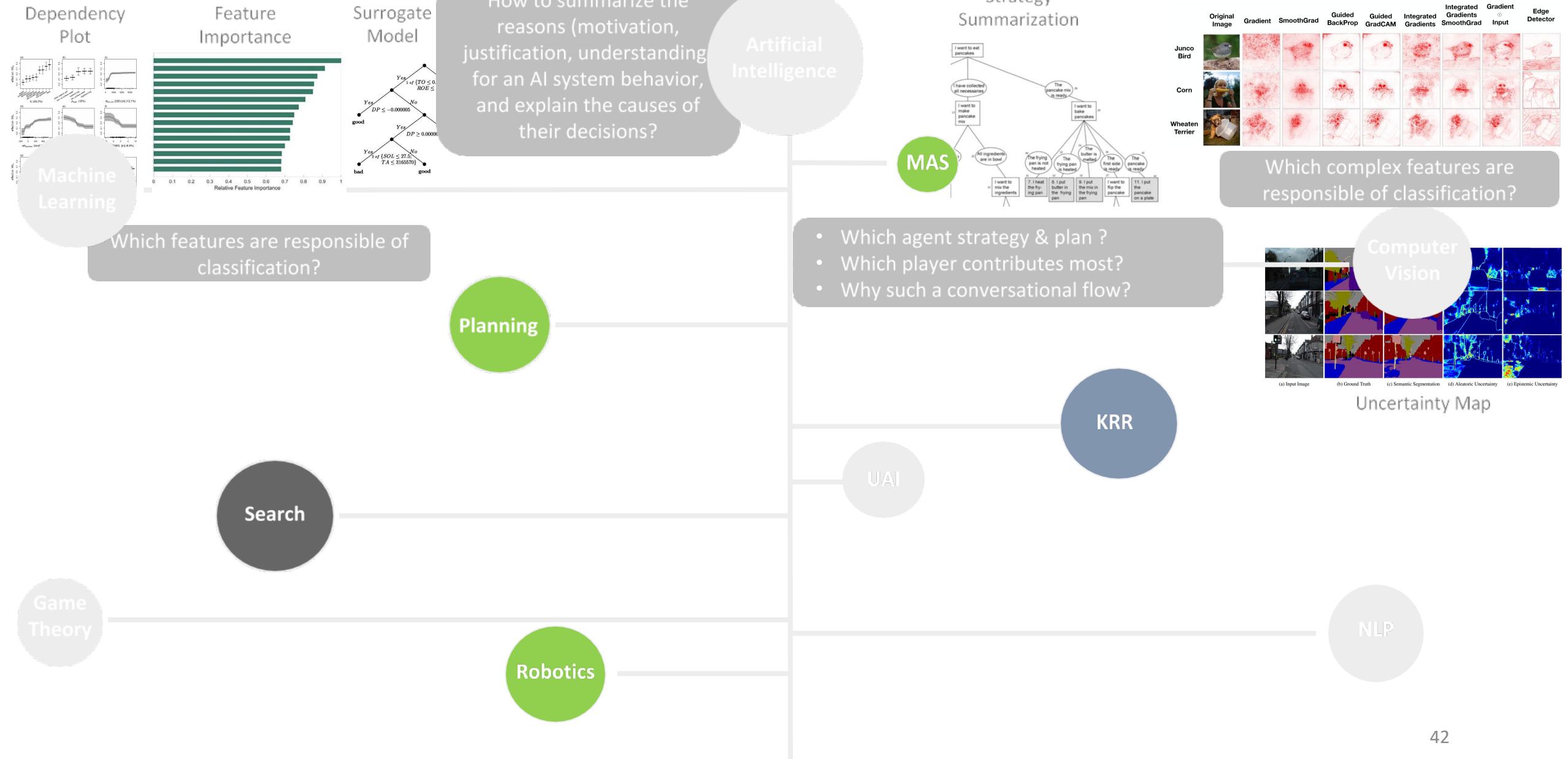
XAI: One Objective, Many ‘AI’s, Many Definitions, Many Approaches



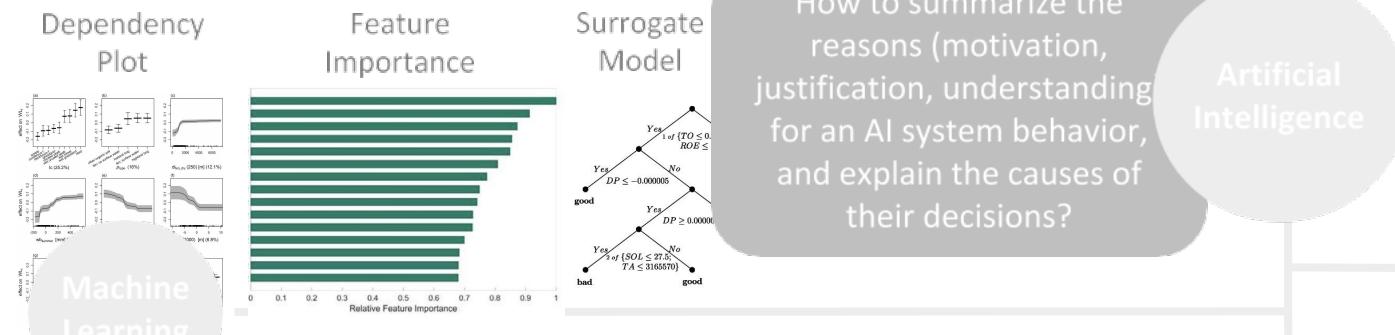
XAI: One Objective, Many ‘AI’s, Many Definitions, Many Approaches



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



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



Which features are responsible of classification?

Plan Refinement

Planning

Which actions are responsible of a plan?

Search

Game Theory

Robotics

Strategy
Summarization

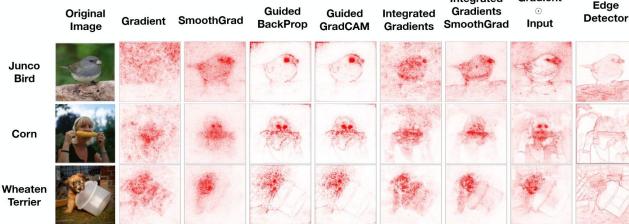
MAS

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

KRR

UAI

Saliency Map



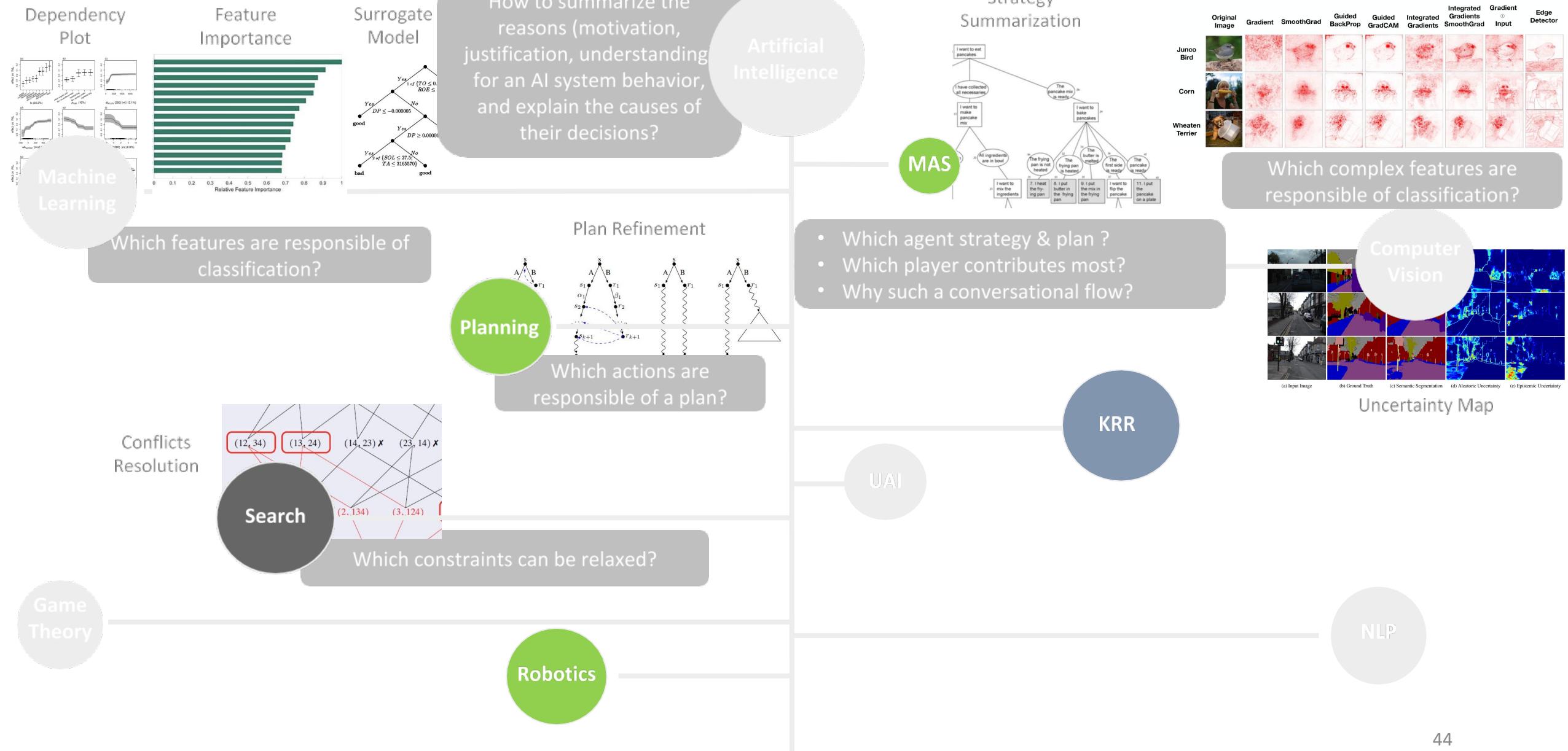
Which complex features are responsible of classification?

Computer Vision

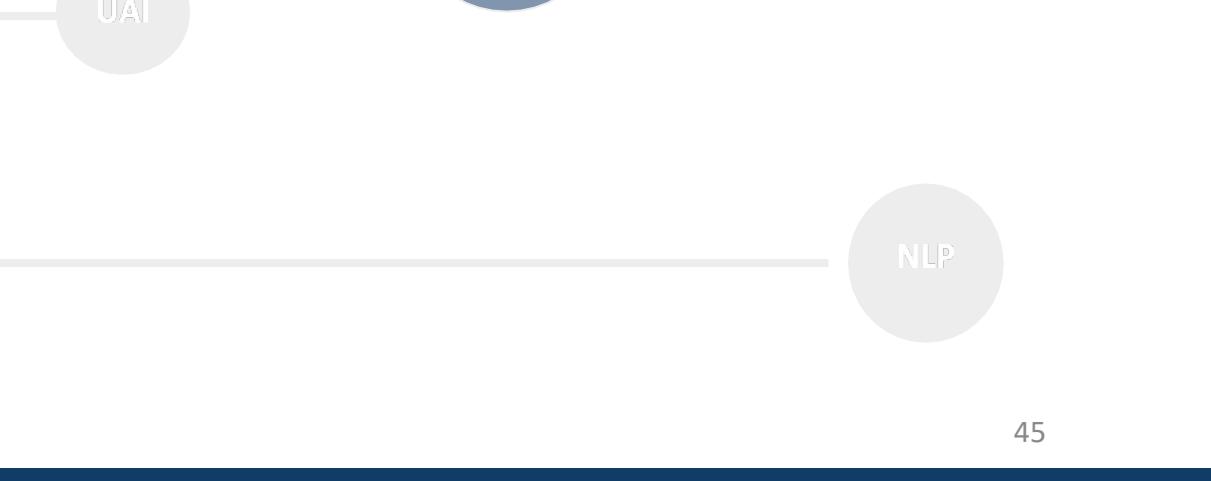
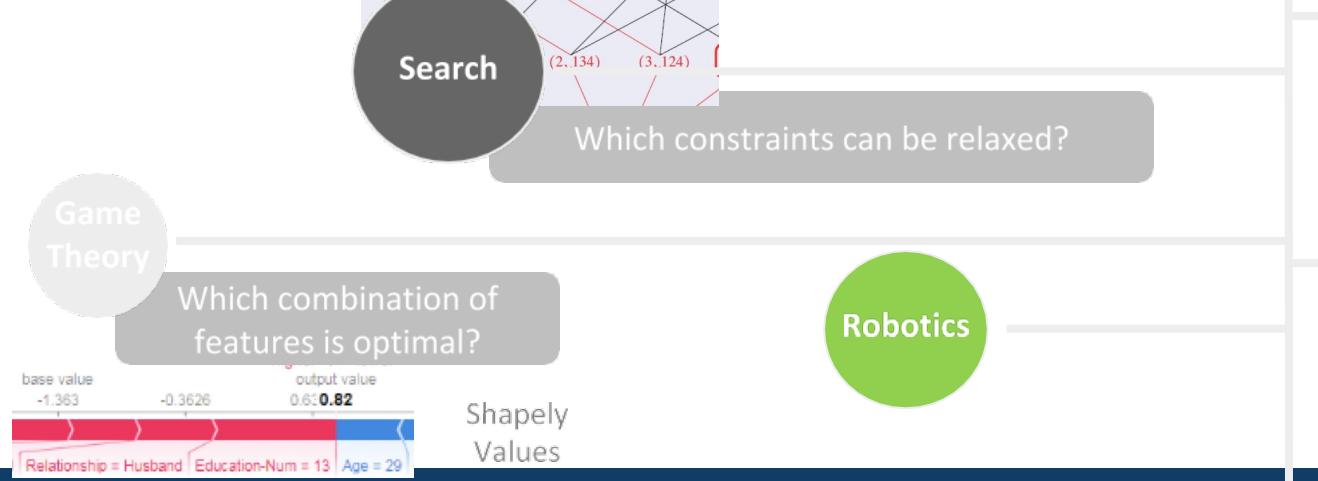
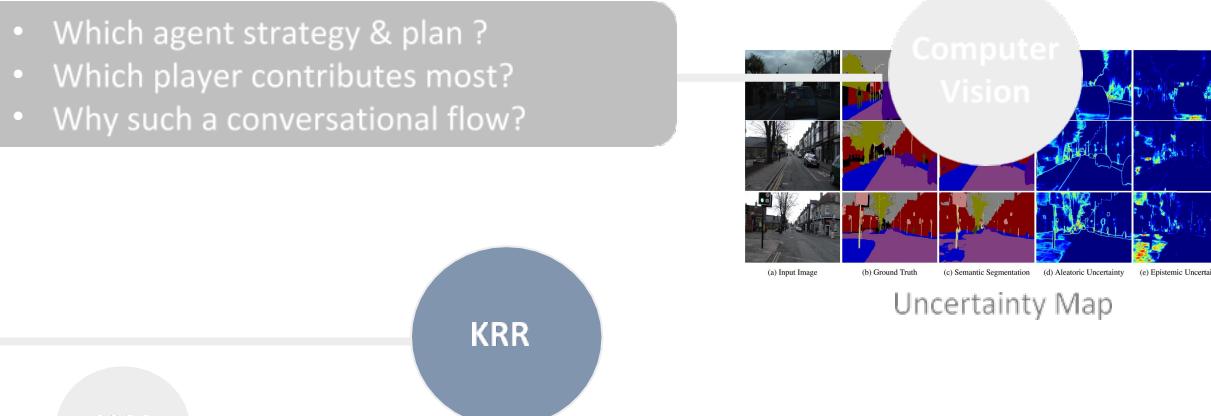
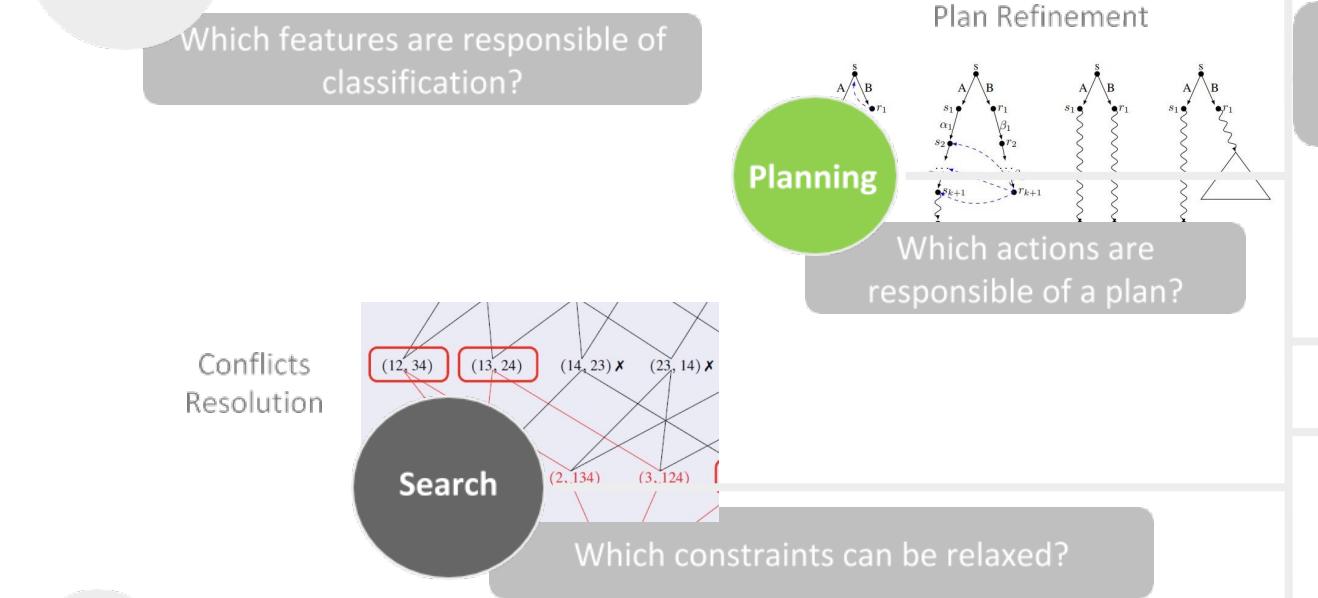
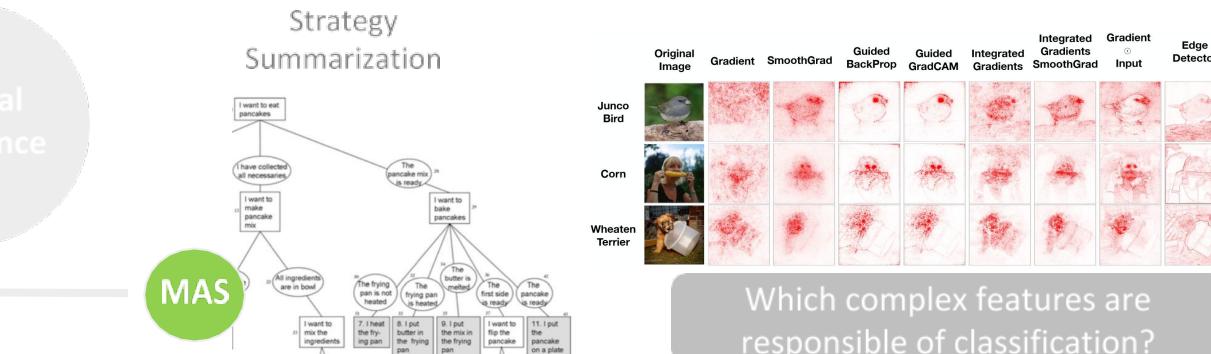
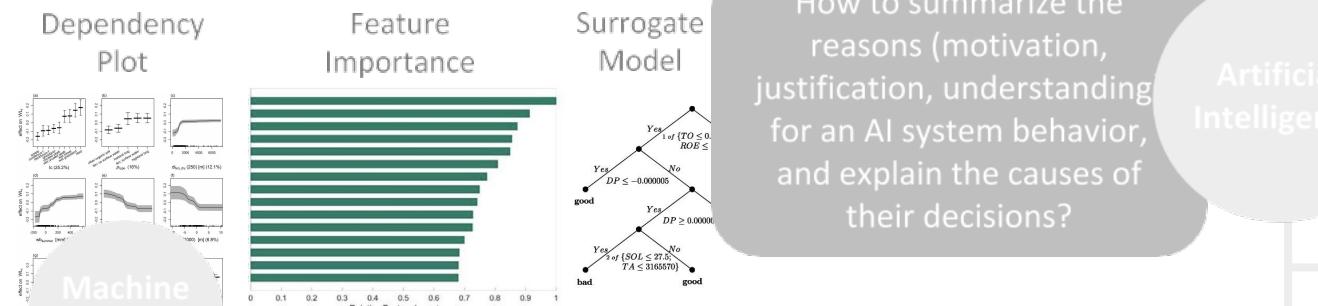


NLP

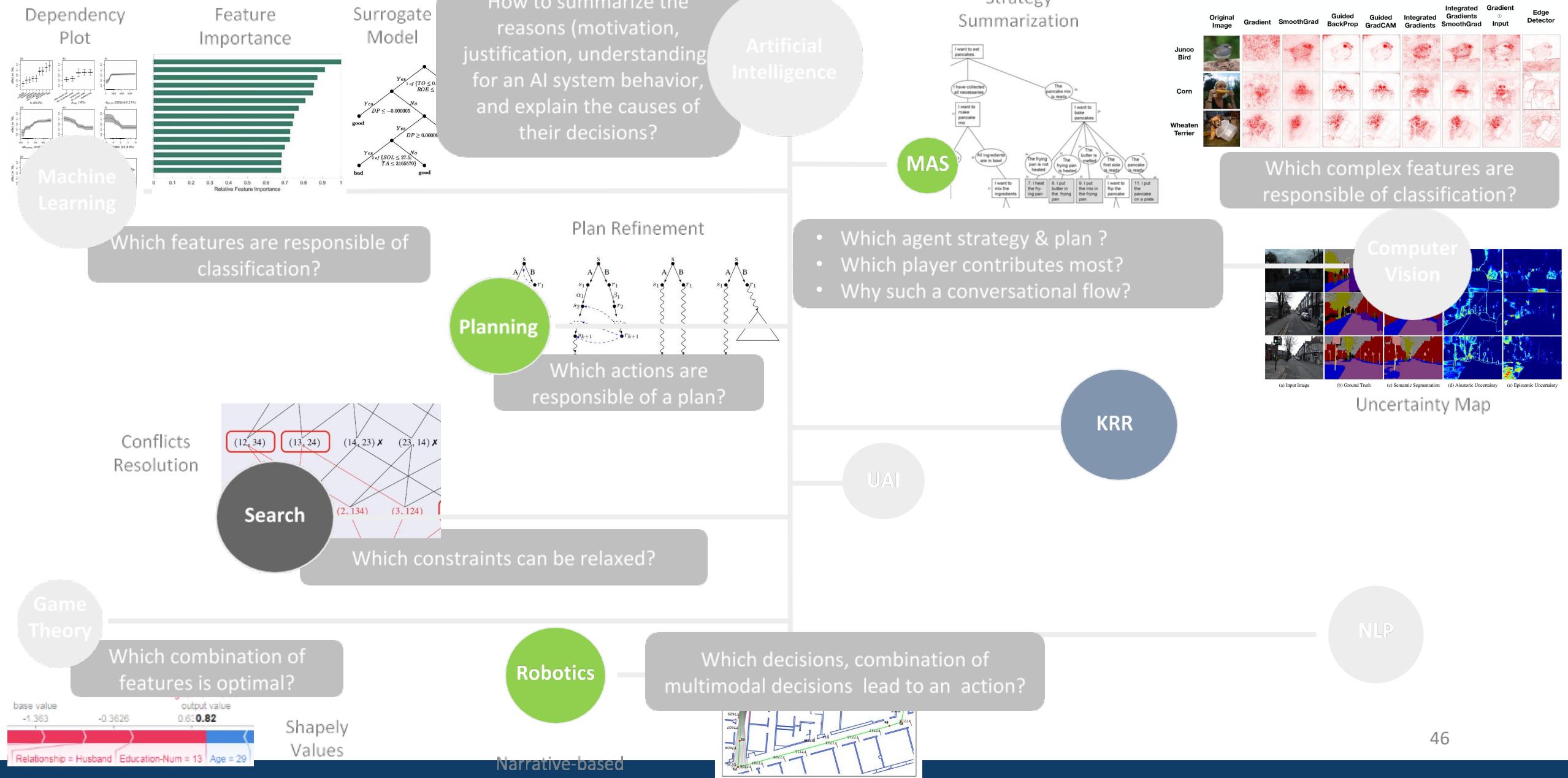
XAI: One Objective, Many ‘AI’s, Many Definitions, Many Approaches



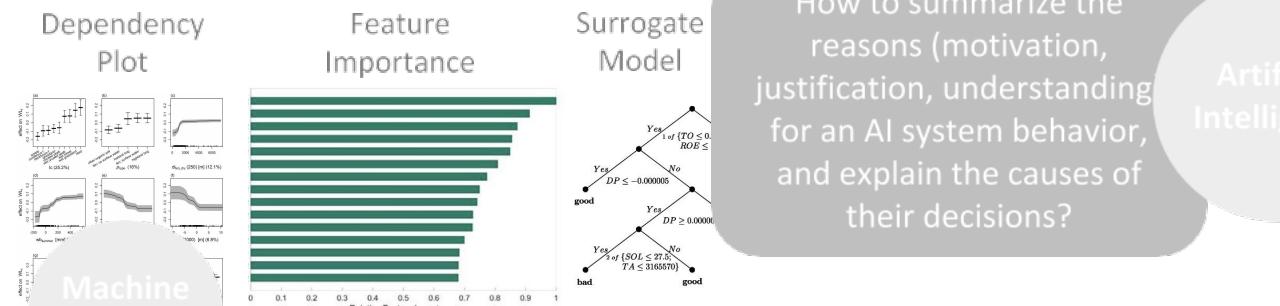
XAI: One Objective, Many ‘AI’s, Many Definitions, Many Approaches



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



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



Which features are responsible of classification?

Conflicts Resolution



Which constraints can be relaxed?

Game Theory

Which combination of features is optimal?

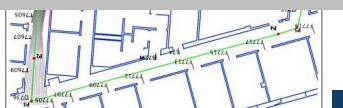


Shapely Values

Robotics

Narrative-based

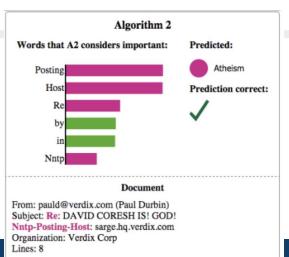
Which decisions, combination of multimodal decisions lead to an action?



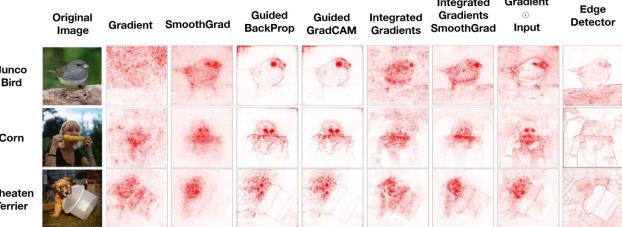
- Which agent strategy & plan ?
- Which player contributes most?
- Why such a conversational flow?

UAI

Machine Learning based



Saliency Map



Which complex features are responsible of classification?

Computer Vision

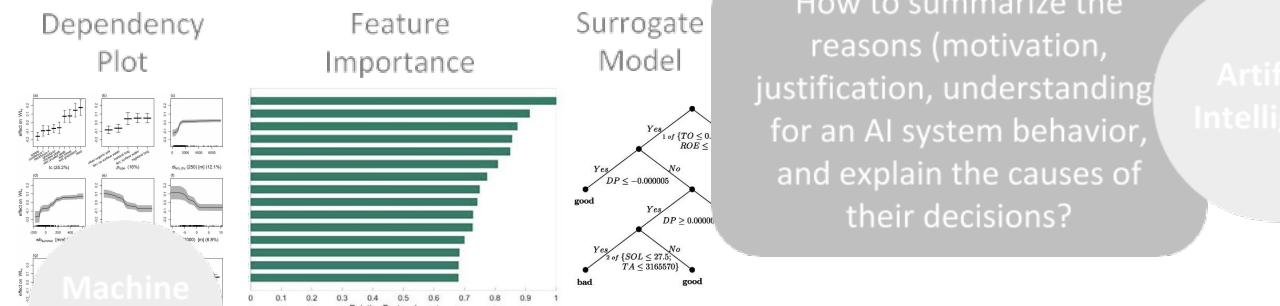


Uncertainty Map

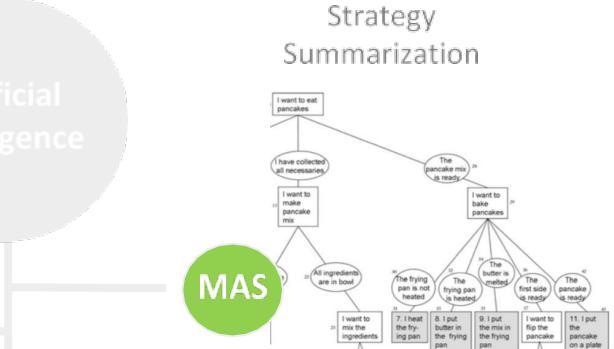
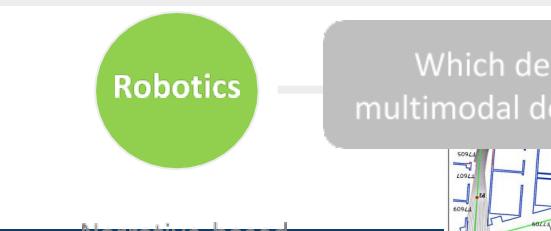
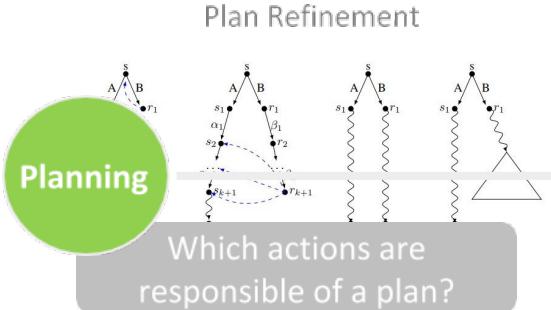
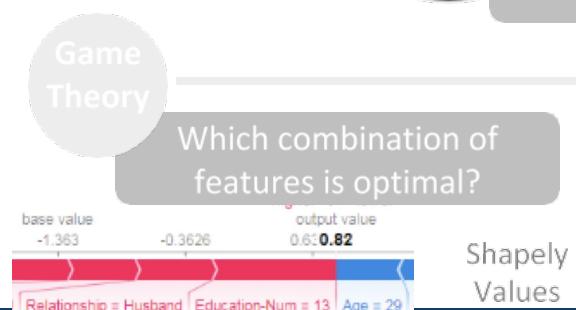
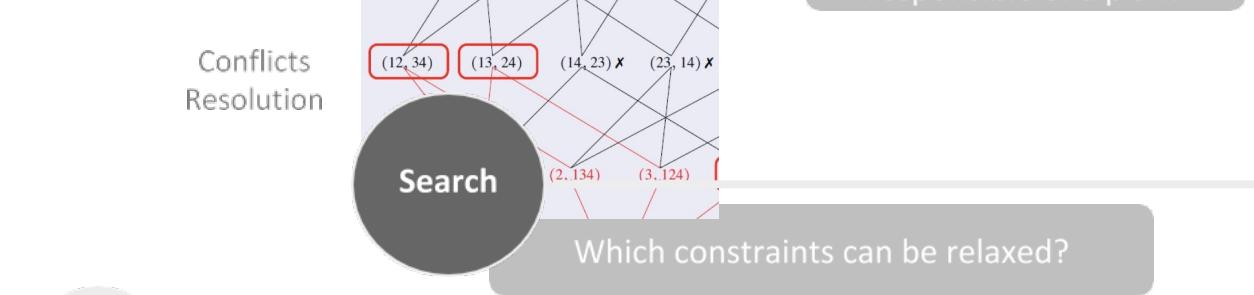
NLP

Which entity is responsible for classification?

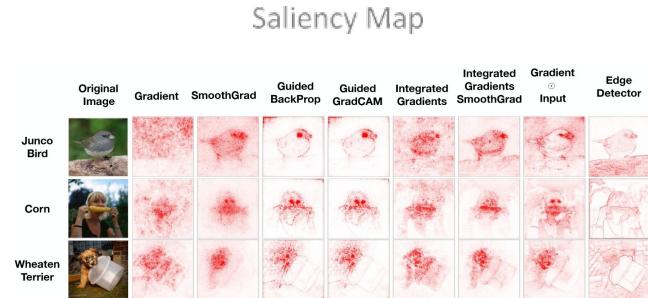
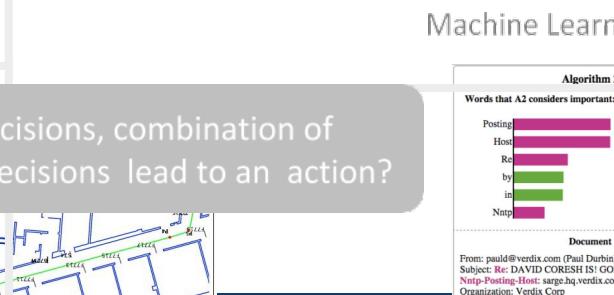
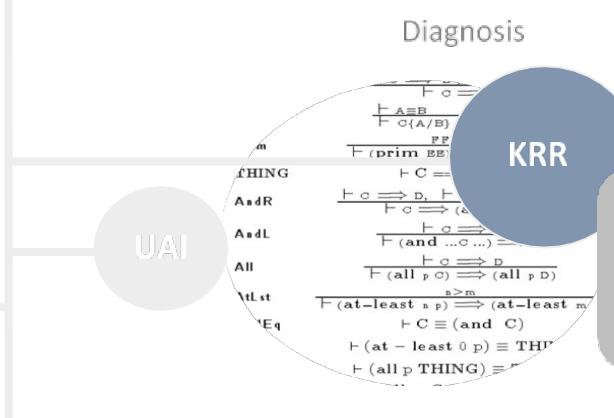
XAI: One Objective, Many ‘AI’s, Many Definitions, Many Approaches



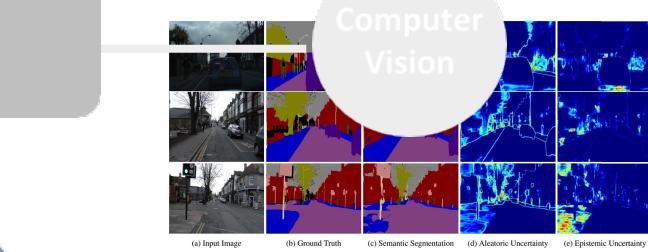
Which features are responsible of classification?



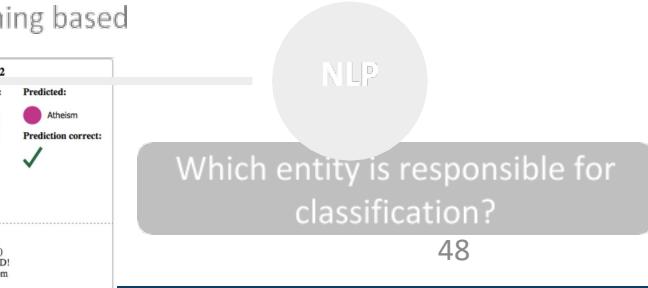
- Which agent strategy & plan ?
- Which player contributes most?
- Why such a conversational flow?



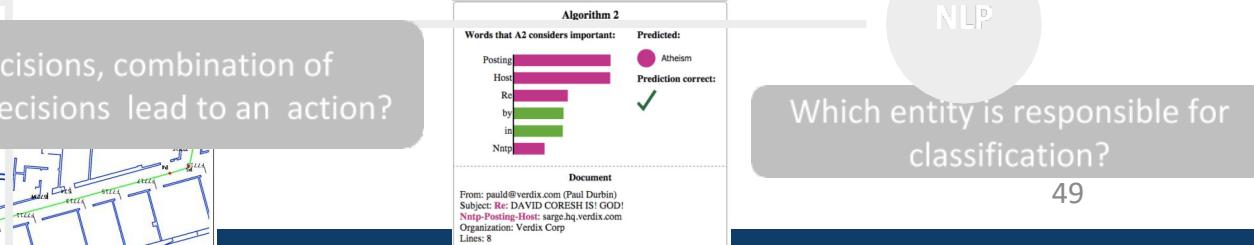
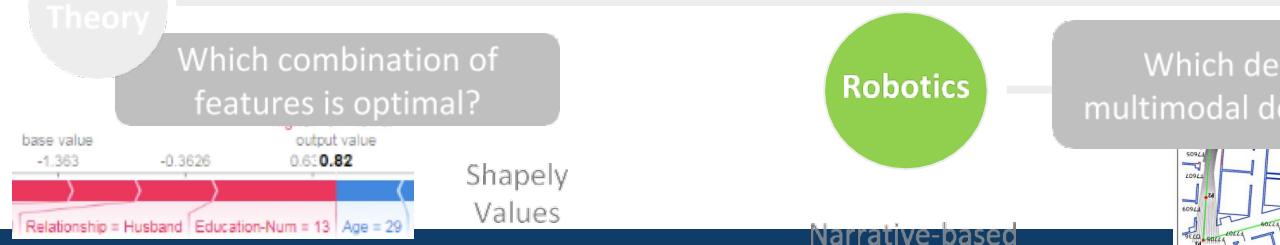
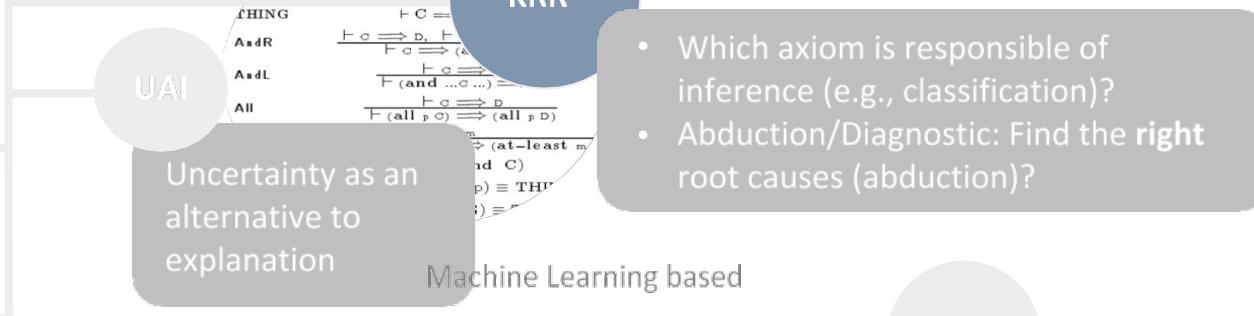
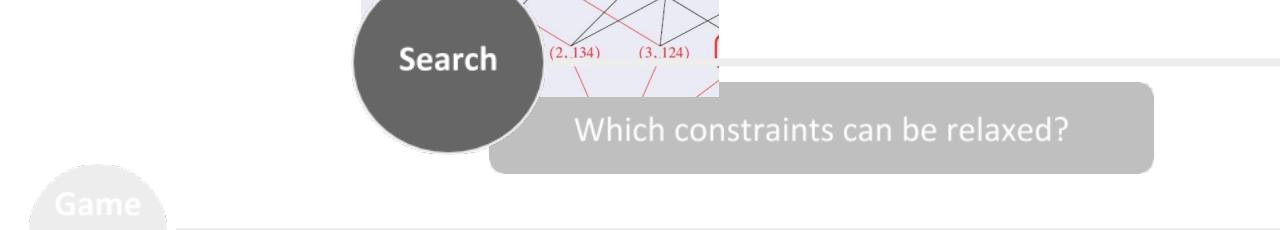
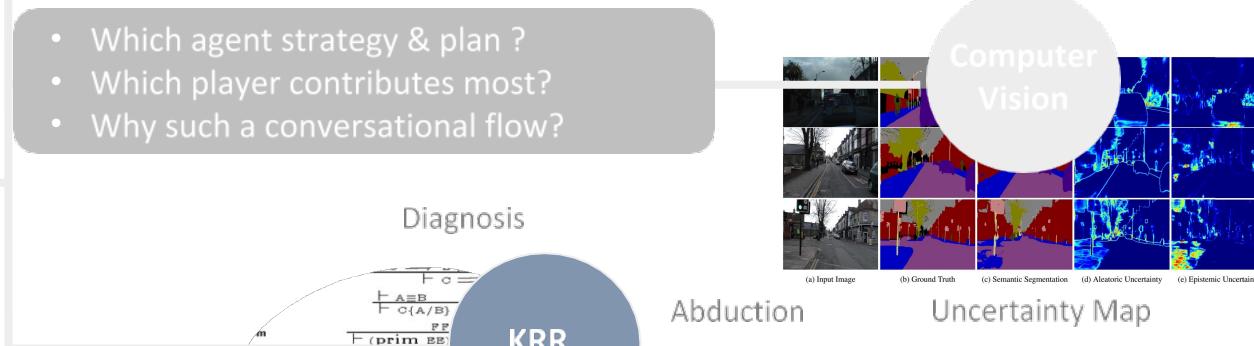
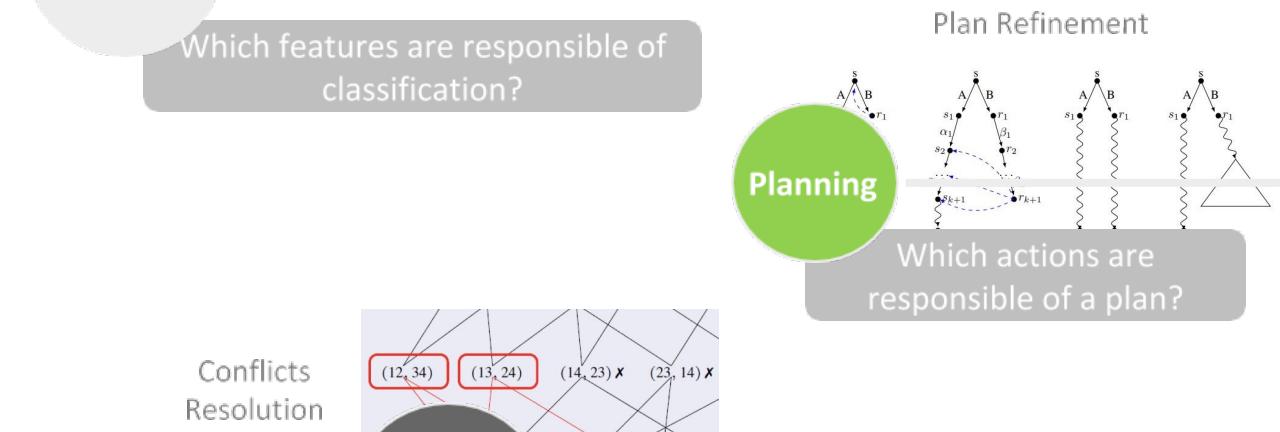
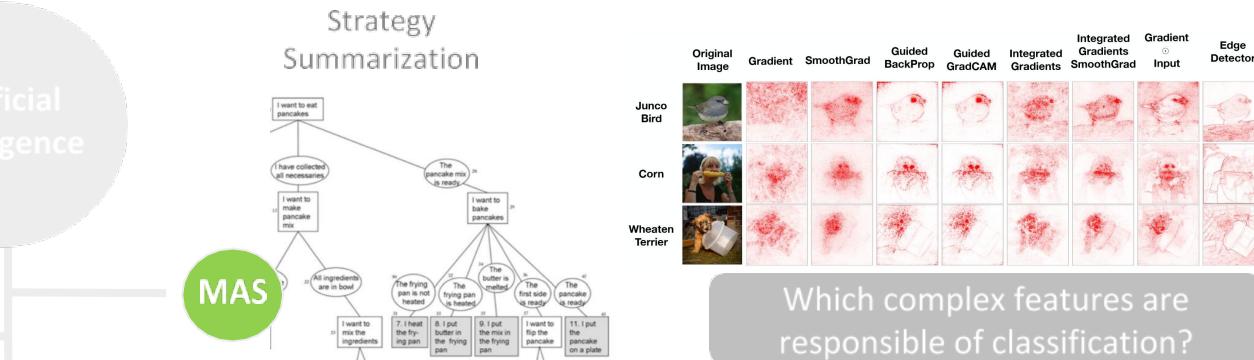
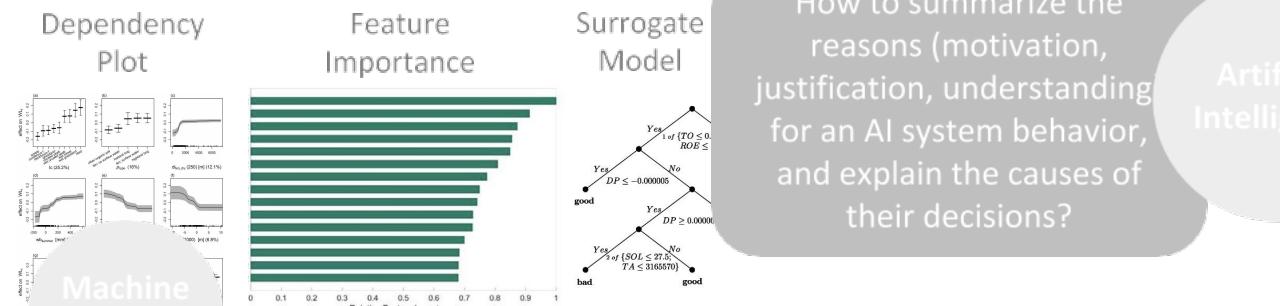
Which complex features are responsible of classification?



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



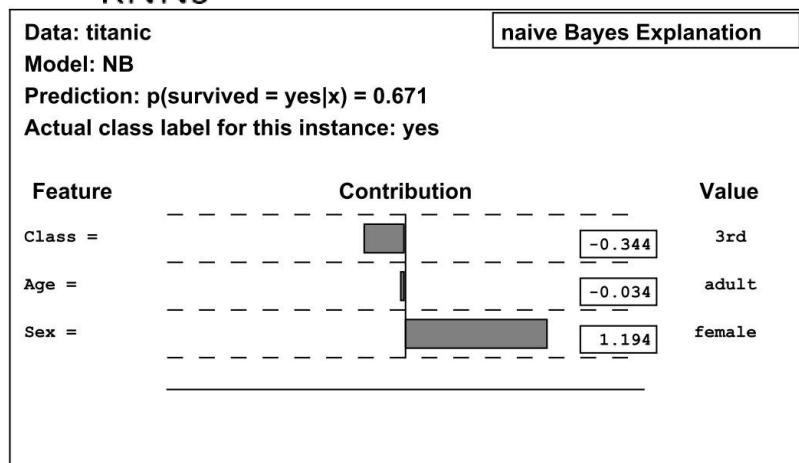
XAI: One Objective, Many ‘AI’s, Many Definitions, Many Approaches



Overview of Explanation in Machine Learning (1)

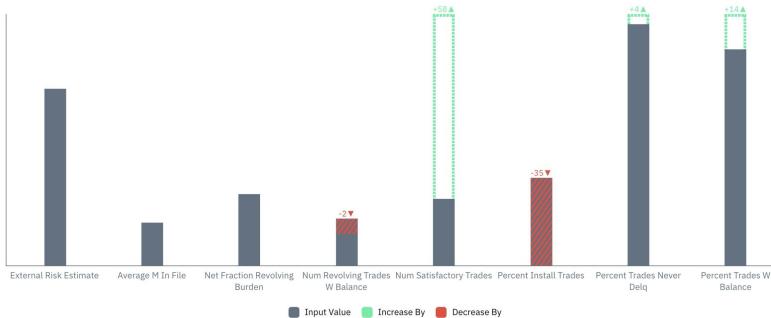
Interpretable Models:

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



Naive Bayes model

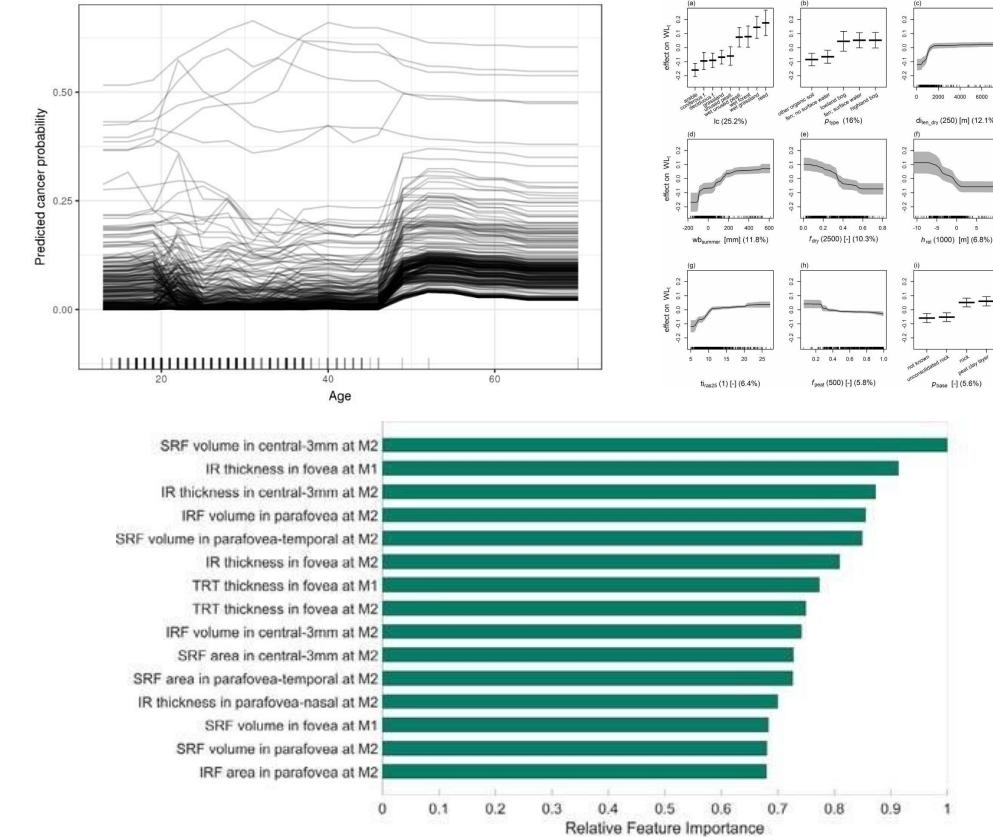
Igor Kononenko. Machine learning for medical diagnosis: history, state of the art and perspective. Artificial Intelligence in Medicine, 23:89–109, 2001.



Counterfactual What-if

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

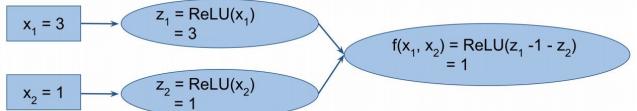
Rory McGrath, Luca Costabello,
Chan Le Van, Paul Sweeney,
Farbod Kamiab, Zhao Shen,
Freddy Lécué: Interpretable Credit
Application Predictions With
Counterfactual Explanations.
CoRR abs/1811.05245 (2018)



(a)
Feature Importance
Partial Dependence Plot
Individual Conditional Expectation
Sensitivity Analysis

Overview of Explanation in Machine Learning (2)

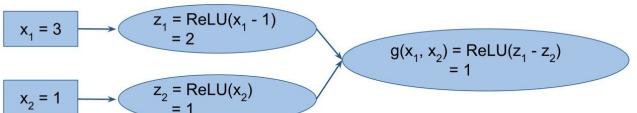
● Artificial Neural Network



Network $f(x_1, x_2)$

Attributions at $x_1 = 3, x_2 = 1$

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



Network $g(x_1, x_2)$

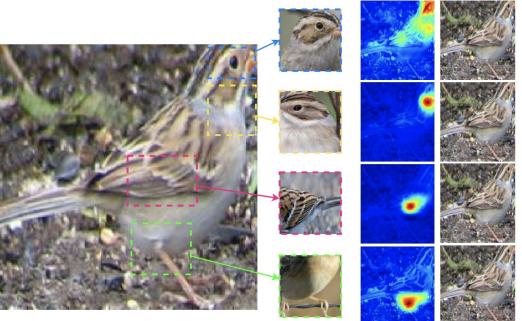
Attributions at $x_1 = 3, x_2 = 1$

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

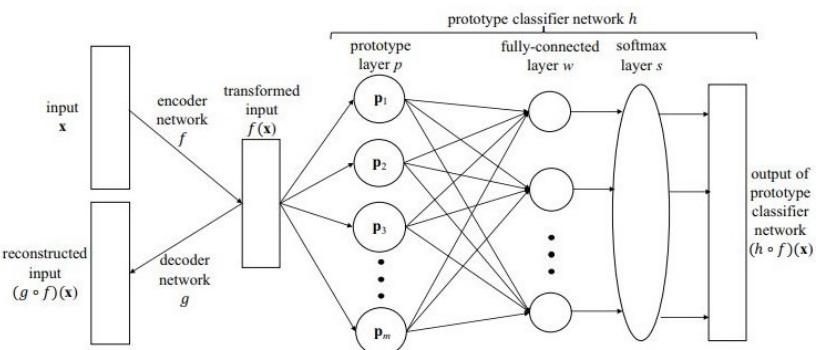
Attribution for Deep Network (Integrated gradient-based)

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

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

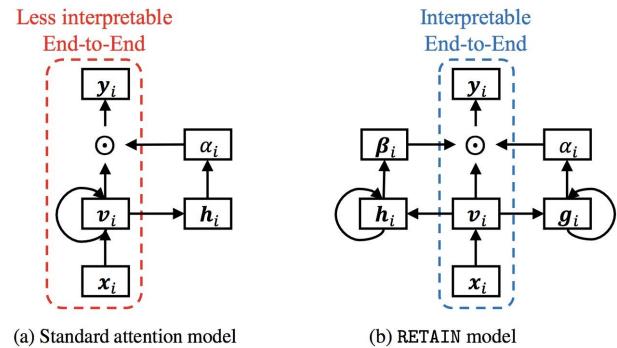


Chaofan Chen, Oscar Li, Alina Barnett, Jonathan Su, Cynthia Rudin: This looks like that: deep learning for interpretable image recognition. CoRR abs/1806.10574 (2018)



Auto-encoder / Prototype

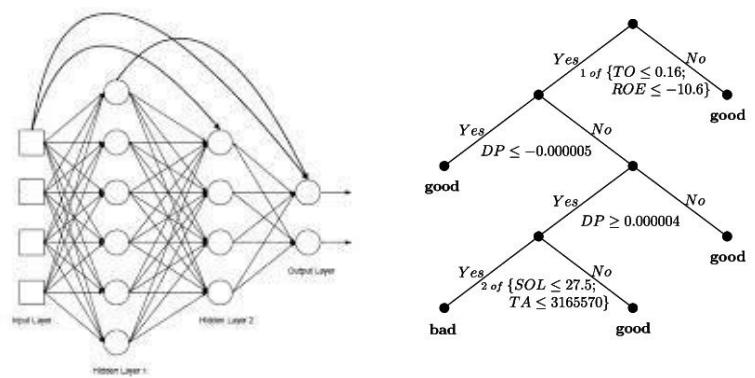
Oscar Li, Hao Liu, Chaofan Chen, Cynthia Rudin: Deep Learning for Case-Based Reasoning Through Prototypes: A Neural Network That Explains Its Predictions. AAAI 2018: 3530-3537



Attention Mechanism

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

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



Surrogate Model

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

Overview of Explanation in Machine Learning (3)

● Computer Vision

Train

res5c unit 924



res5c unit 2001



inception_5b unit 626

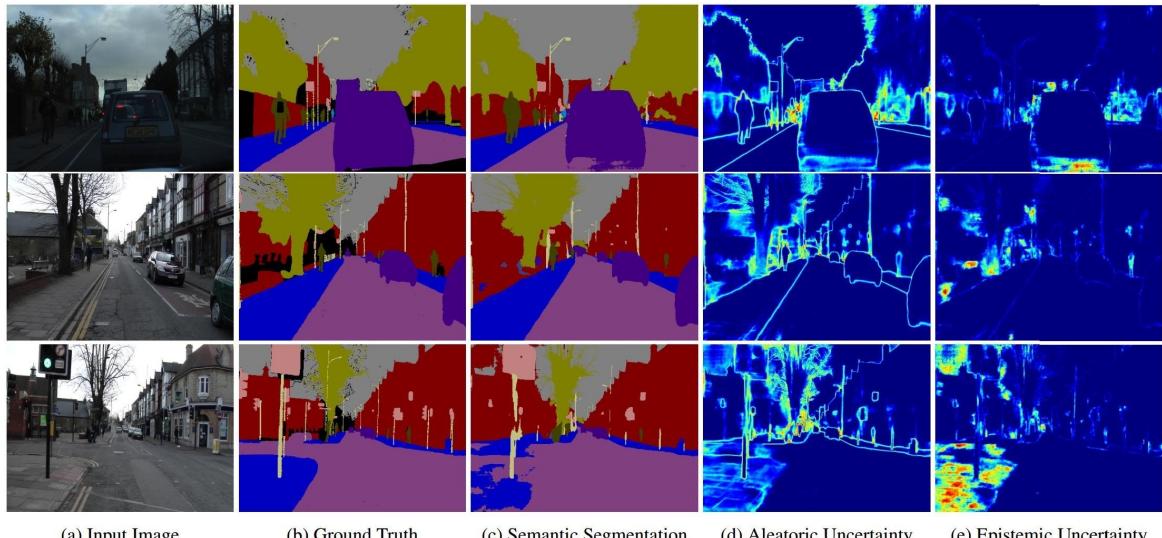


inception_5b unit 415



Interpretable Units

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



Uncertainty Map

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

Airplane

res5c unit 1243



res5c unit 1379



inception_4e unit 92



Western Grebe

Description: This is a large bird with a white neck and a black back in the water.

Class Definition: The Western Grebe is a waterbird with a yellow pointy beak, white neck and belly, and black back.

Explanation: This is a *Western Grebe* because this bird has a long white neck, pointy yellow beak and red eye.

Laysan Albatross

Description: This is a large flying bird with black wings and a white belly.

Class Definition: The Laysan Albatross is a large seabird with a hooked yellow beak, black back and white belly.

Visual Explanation: This is a *Laysan Albatross* because this bird has a large wingspan, hooked yellow beak, and white belly.

Laysan Albatross

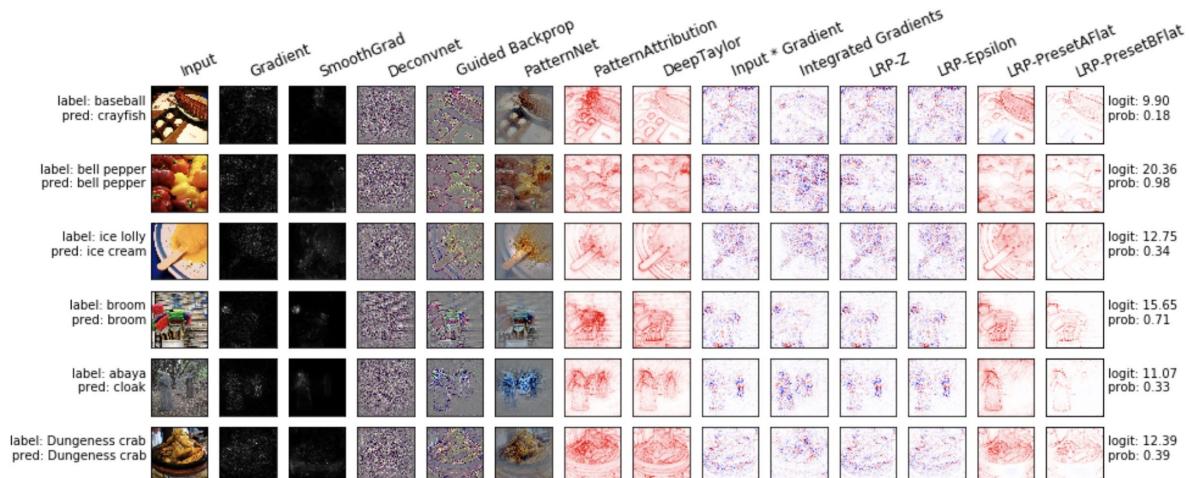
Description: This is a large bird with a white neck and a black back in the water.

Class Definition: The Laysan Albatross is a large seabird with a hooked yellow beak, black back and white belly.

Visual Explanation: This is a *Laysan Albatross* because this bird has a hooked yellow beak white neck and black back.

Visual Explanation

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

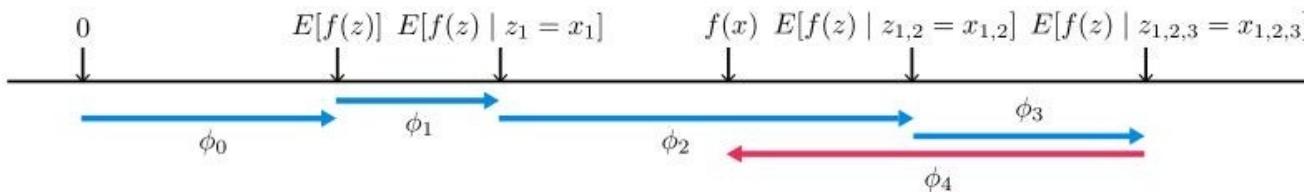
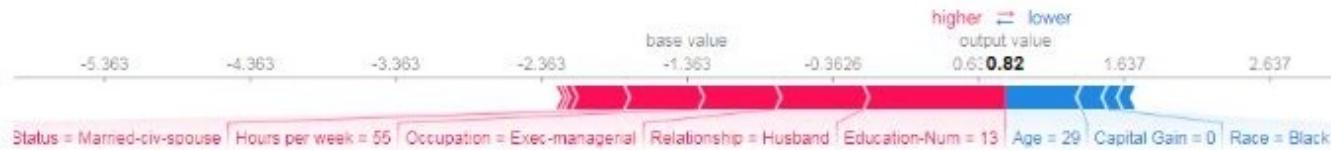


Saliency Map

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

Overview of Explanation in Different AI Fields (1)

● Game Theory

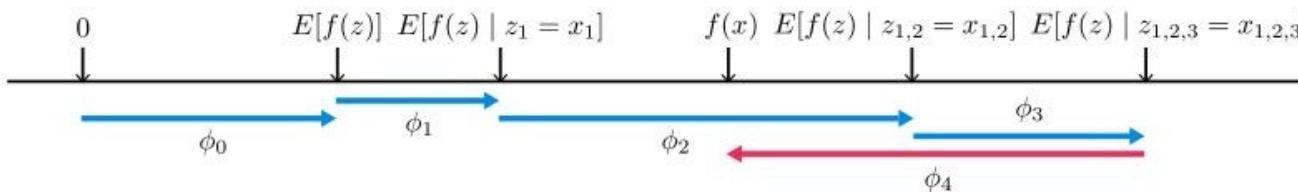
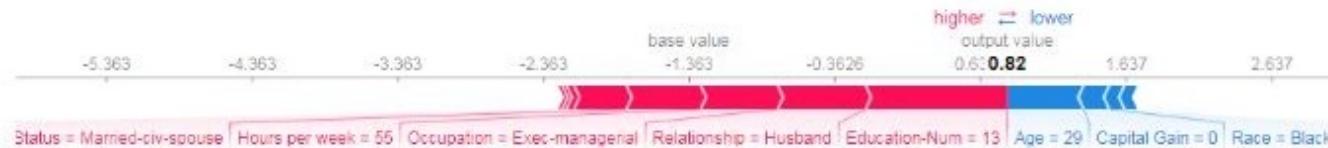


Shapley Additive Explanation

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

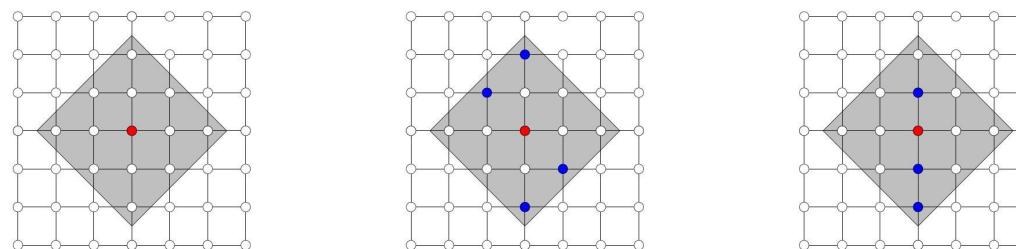
Overview of Explanation in Different AI Fields (1)

● Game Theory



Shapley Additive Explanation

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



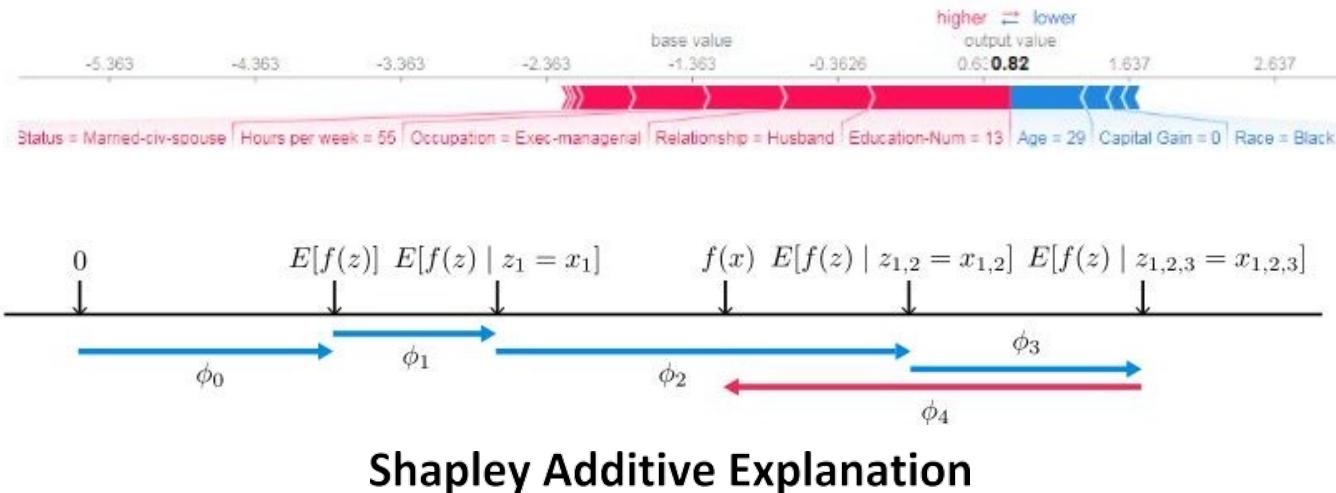
Local Shapley
Connected Shapley

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

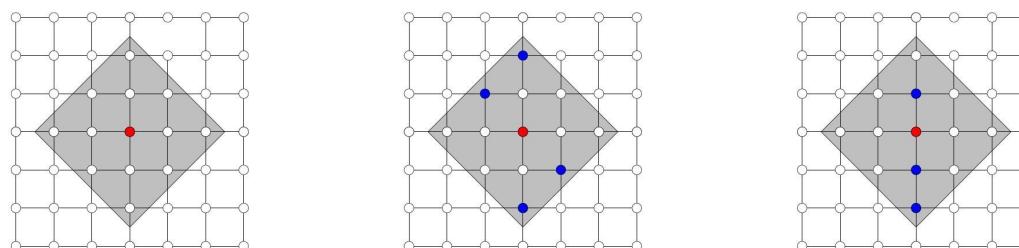


Scott M. Lundberg, Su-In Lee: A Unified Approach to Interpreting Model Predictions. NIPS 2017:
4768-4777

**instance-wise 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.

Local Shapley
Connected Shapley



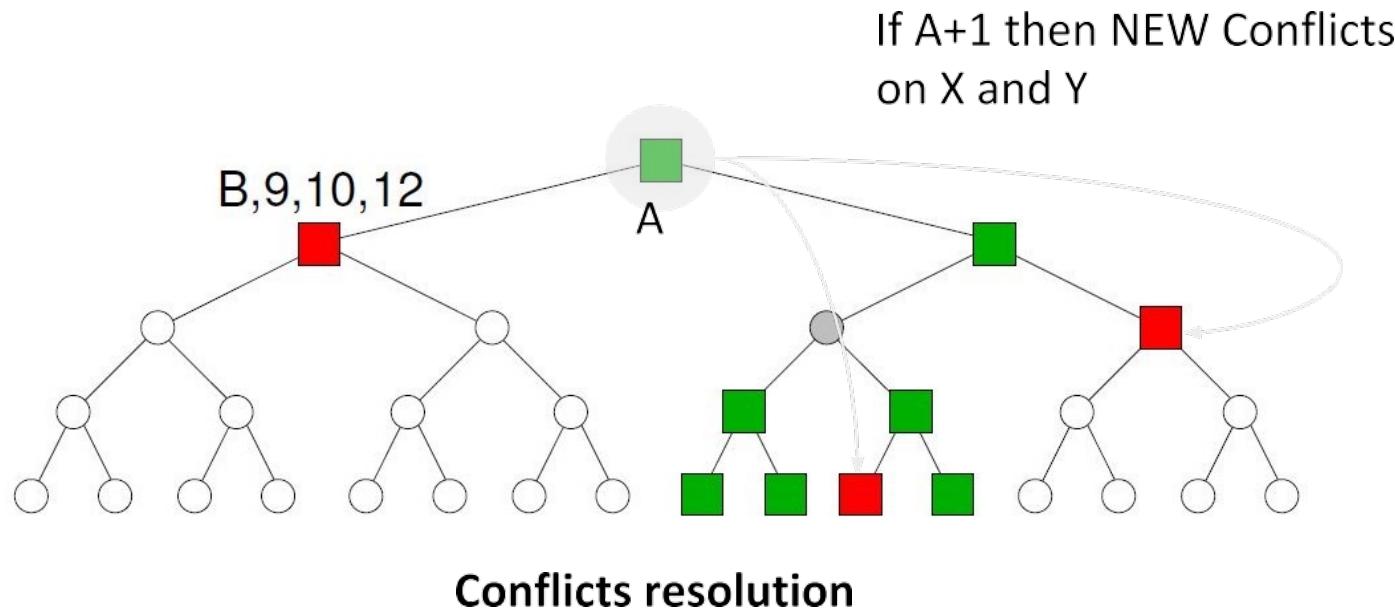
L-Shapley and C-Shapley (with graph structure)

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

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

Overview of Explanation in Different AI Fields (2)

- Search and Constraint Satisfaction



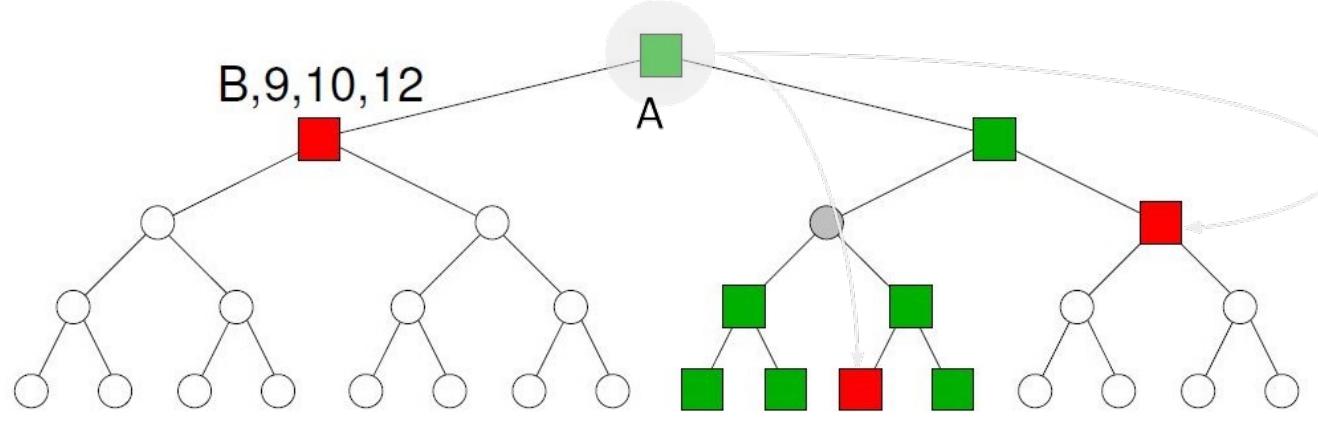
Barry O'Sullivan, Alexandre Papadopoulos, Boi Faltings, Pearl Pu: Representative Explanations for Over-Constrained Problems. AAAI 2007: 323-328

Robustness Computation

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

Overview of Explanation in Different AI Fields (2)

- Search and Constraint Satisfaction



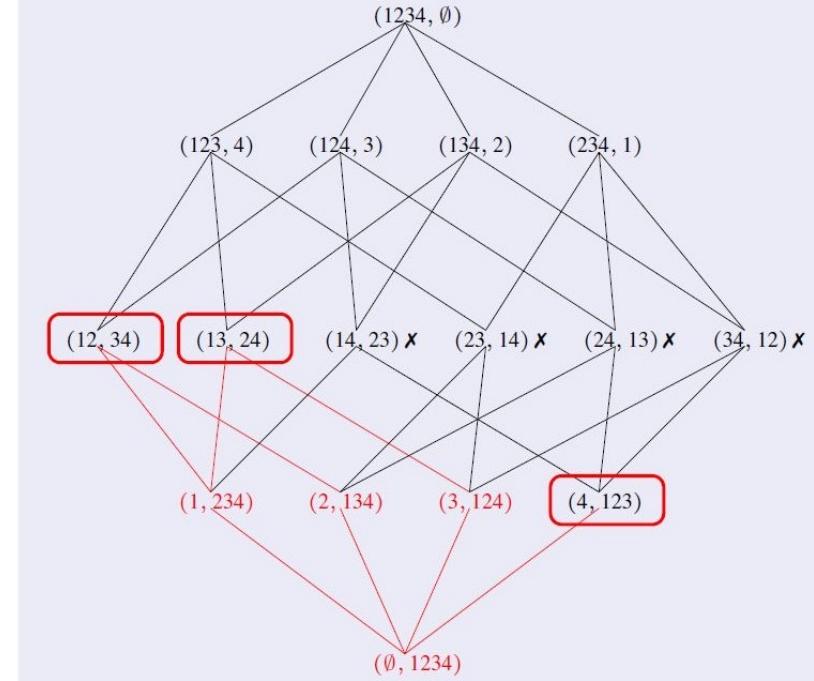
Conflicts resolution

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

Robustness Computation

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

Explanations



Constraints relaxation

Ulrich Junker: QUICKPLAIN: 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	$\vdash C \Rightarrow C$	
Trans	$\frac{\vdash c \Rightarrow d, \vdash d \Rightarrow e}{\vdash c \Rightarrow e}$	
Eq	$\frac{\vdash A \equiv B \quad \vdash c \Rightarrow D}{\vdash C\{A/B\} \Rightarrow D\{A/B\}}$	
Prim	$\frac{FF \in EE}{\vdash (\text{prim } EE) \Rightarrow (\text{prim } FF)}$	
THING	$\vdash C \Rightarrow \text{THING}$	
AndR	$\frac{\vdash c \Rightarrow d, \vdash c \Rightarrow (\text{and } EE)}{\vdash c \Rightarrow (\text{and } d \text{ } EE)}$	
AndL	$\frac{\vdash c \Rightarrow E}{\vdash (\text{and } \dots c \dots) \Rightarrow E}$	
All	$\frac{\vdash c \Rightarrow D}{\vdash (\text{all } p \text{ } C) \Rightarrow (\text{all } p \text{ } D)}$	
AtLst	$\frac{}{\vdash (\text{at-least } n \text{ } p) \Rightarrow (\text{at-least } m \text{ } p)}$	
AndEq	$\vdash C \equiv (\text{and } C)$	
AtLs0	$\vdash (\text{at least } 0 \text{ } p) \equiv \text{THING}$	
All-thing	$\vdash (\text{all } p \text{ } \text{THING}) \equiv \text{THING}$	
All-and	$\vdash (\text{and } (\text{all } p \text{ } C) (\text{all } p \text{ } D) \dots) \equiv (\text{and } (\text{all } p \text{ } (\text{and } C \text{ } D)) \dots)$	

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

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

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

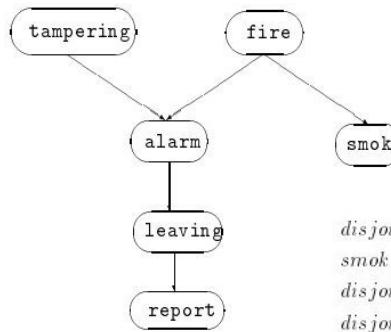
Overview of Explanation in Different AI Fields (3)

• Knowledge Representation and Reasoning

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

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

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



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

$\text{disjoint}([\text{fire}(yes) : 0.01, \text{fire}(no) : 0.99]).$
 $\text{smoke}(Sm) \leftarrow \text{fire}(Fi) \wedge c_smoke(Sm, Fi).$
 $\text{disjoint}([c_smoke(yes, yes) : 0.9, c_smoke(no, yes) : 0.1]).$
 $\text{disjoint}([c_smoke(yes, no) : 0.01, c_smoke(no, no) : 0.99]).$

Abduction Reasoning (in Bayesian Network)

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

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

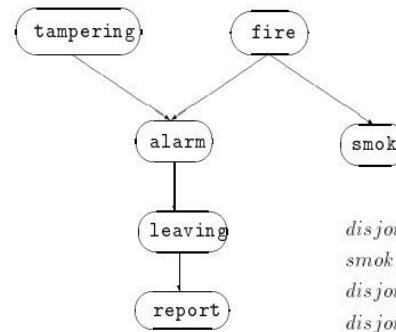
Ref	$\vdash C \Rightarrow C$
Trans	$\vdash c \Rightarrow d, \vdash d \Rightarrow e \quad \vdash c \Rightarrow e$
Eq	$\vdash A \equiv B \quad \vdash c \Rightarrow D \quad \vdash c(A/B) \Rightarrow D(A/B)$
Prim	$\vdash (\text{prim } EE) \Rightarrow (\text{prim } FF)$
THING	$\vdash C \Rightarrow \text{THING}$
AndR	$\vdash c \Rightarrow d, \vdash c \Rightarrow (\text{and } EE) \quad \vdash c \Rightarrow (\text{and } D \text{ EE})$
AndL	$\vdash c \Rightarrow e \quad \vdash (\text{and } ...c ...) \Rightarrow e$
All	$\vdash c \Rightarrow d \quad \vdash (\text{all } p \ C) \Rightarrow (\text{all } p \ D)$
AtLst	$\vdash (\text{at-least } n \ p) \Rightarrow (\text{at-least } m \ p)$
AndEq	$\vdash C \equiv (\text{and } C)$
AtLs0	$\vdash (\text{at - least } 0 \ p) \equiv \text{THING}$
All-thing	$\vdash (\text{all } p \ \text{THING}) \equiv \text{THING}$
All-and	$\vdash (\text{and } (\text{all } p \ C) (\text{all } p \ D) ...) \equiv (\text{and } (\text{all } p \ (\text{and } C \ D)) ...)$

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

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

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

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

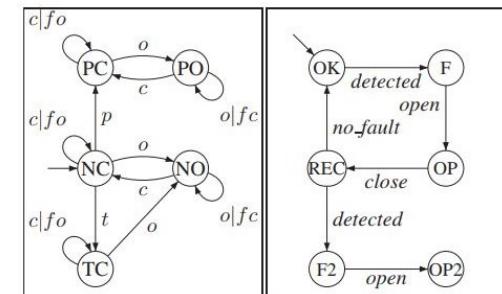


$P(\text{alarm}|\text{fire} \wedge \neg \text{tampering}) = 0.99$
 $P(\text{alarm}|\neg \text{fire} \wedge \text{tampering}) = 0.85$
 $P(\text{alarm}|\neg \text{fire} \wedge \neg \text{tampering}) = 0.0001$
 $P(\text{leaving}|\text{alarm}) = 0.88$
 $P(\text{leaving}|\neg \text{alarm}) = 0.001$
 $P(\text{report}|\text{leaving})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

$\text{disjoint}([\text{fire}(yes) : 0.01, \text{fire}(no) : 0.99]).$
 $\text{smoke}(Sm) \leftarrow \text{fire}(Fi) \wedge c_smoke(Sm, Fi).$
 $\text{disjoint}([c_smoke(yes, yes) : 0.9, c_smoke(no, yes) : 0.1]).$
 $\text{disjoint}([c_smoke(yes, no) : 0.01, c_smoke(no, no) : 0.99]).$

Abduction Reasoning (in Bayesian Network)

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



Diagnosis Inference

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

Overview of Explanation in Different AI Fields (4)

• Multi-agent Systems

MAS INFRASTRUCTURE	INDIVIDUAL AGENT INFRASTRUCTURE
MAS INTEROPERATION Translation Services Interoperation Services	INTEROPERATION Interoperation Modules
CAPABILITY TO AGENT MAPPING Middle Agents	CAPABILITY TO AGENT MAPPING Middle Agents Components
NAME TO LOCATION MAPPING ANS	NAME TO LOCATION MAPPING ANS Component
SECURITY Certificate Authority Cryptographic Services	SECURITY Security Module private/public Keys
PERFORMANCE SERVICES MAS Monitoring Reputation Services	PERFORMANCE SERVICES Performance Services Modules
MULTIAGENT MANAGEMENT SERVICES Logging, Activity 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 Transfer 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. Giampa: The RETSINA MAS Infrastructure. Autonomous Agents and Multi-Agent Systems 7(1-2): 29-48 (2003)

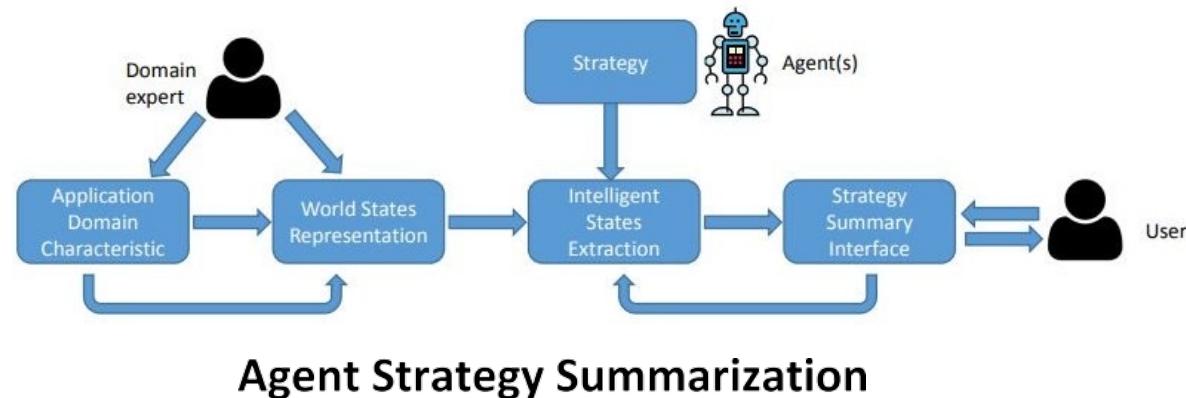
Overview of Explanation in Different AI Fields (4)

• Multi-agent Systems

MAS INFRASTRUCTURE	INDIVIDUAL AGENT INFRASTRUCTURE
MAS INTEROPERATION Translation Services Interoperation Services	INTEROPERATION Interoperation Modules
CAPABILITY TO AGENT MAPPING Middle Agents	CAPABILITY TO AGENT MAPPING Middle Agents Components
NAME TO LOCATION MAPPING ANS	NAME TO LOCATION MAPPING ANS Component
SECURITY Certificate Authority Cryptographic Services	SECURITY Security Module private/public Keys
PERFORMANCE SERVICES MAS Monitoring Reputation Services	PERFORMANCE SERVICES Performance Services Modules
MULTIAGENT MANAGEMENT SERVICES Logging, Activity 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 Transfer 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. Giampa: The RETSINA MAS Infrastructure. Autonomous Agents and Multi-Agent Systems 7(1-2): 29-48 (2003)



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

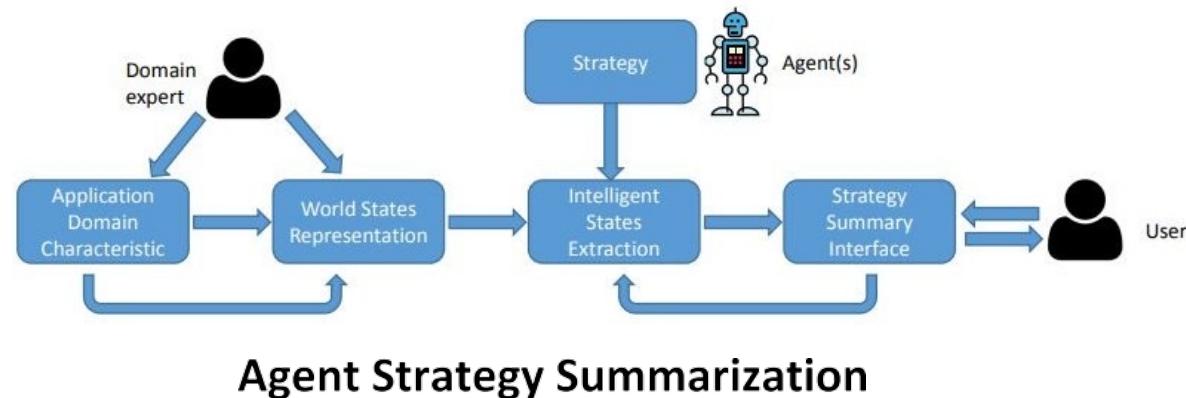
Overview of Explanation in Different AI Fields (4)

• Multi-agent Systems

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

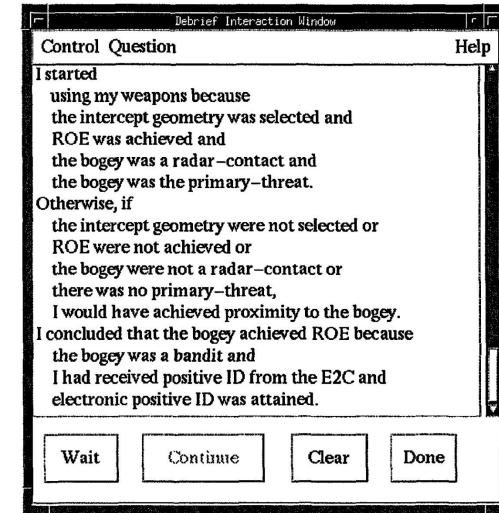
Explanation of Agent Conflicts & Harmful Interactions

Katia P. Sycara, Massimo Paolucci, Martin Van Velsen, Joseph A. Giampa: 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



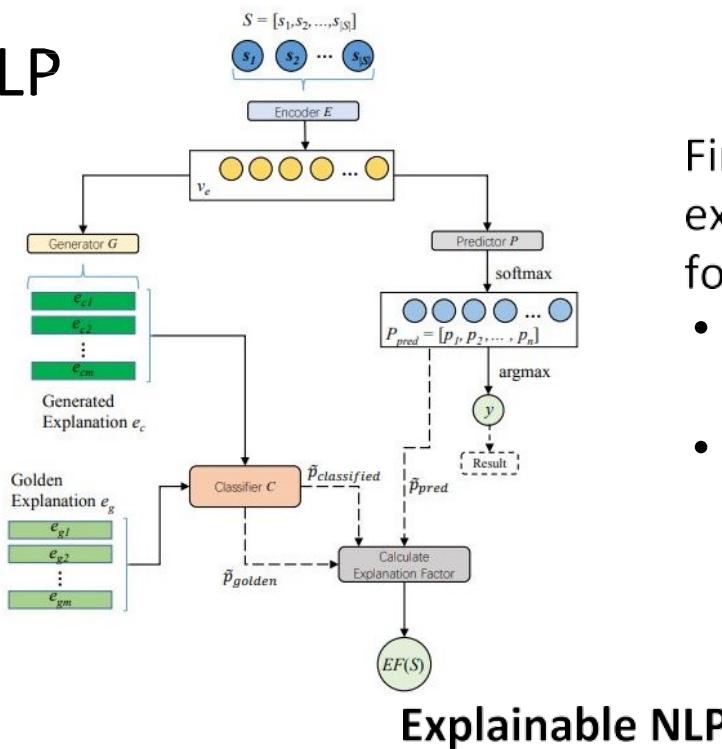
Explainable Agents

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



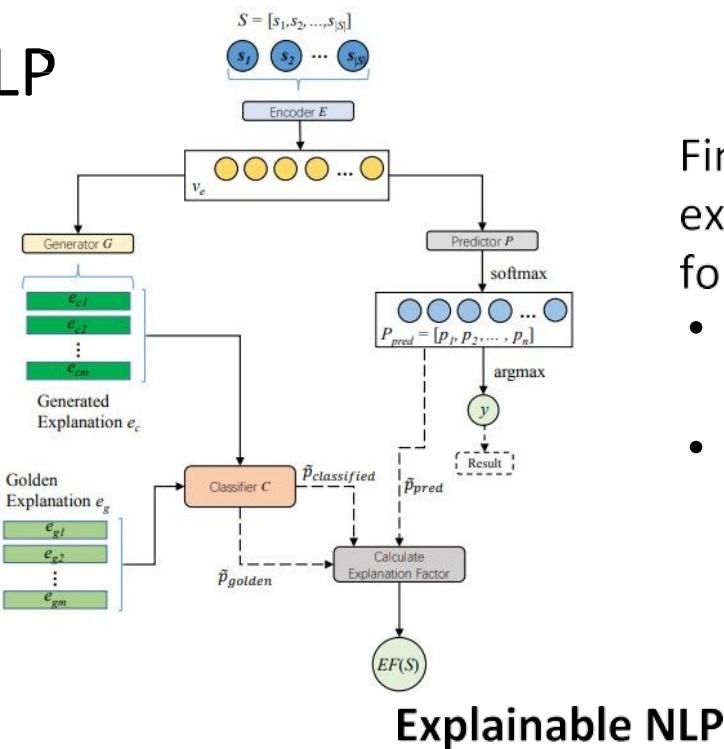
Fine-grained explanations are in the form of:

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

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

Overview of Explanation in Different AI Fields (5)

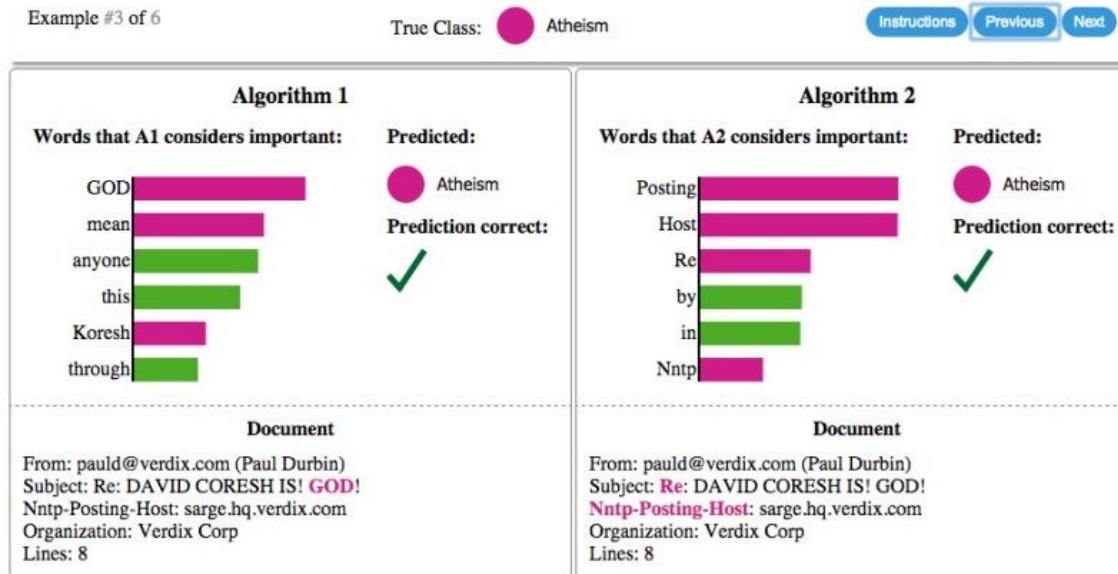
- NLP



Explainable NLP

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

- Fine-grained explanations are in the form of:
- texts in a real-world dataset;
 - Numerical scores

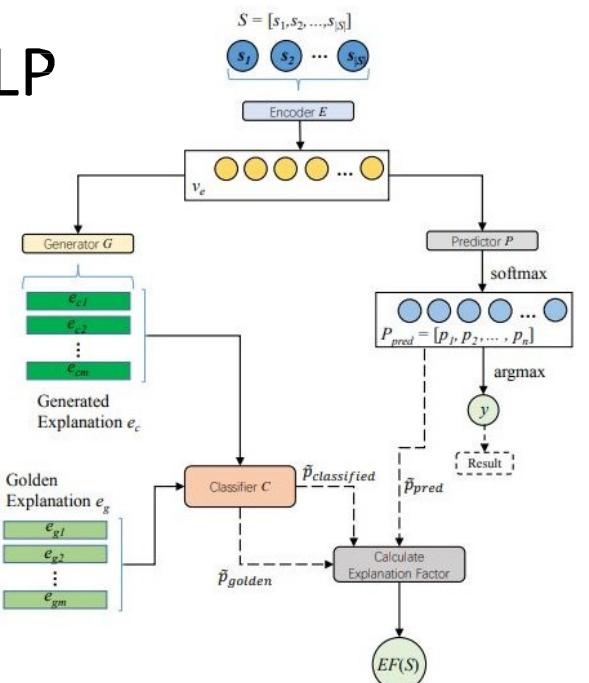


LIME for NLP

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

Overview of Explanation in Different AI Fields (5)

• NLP



Explainable NLP

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

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

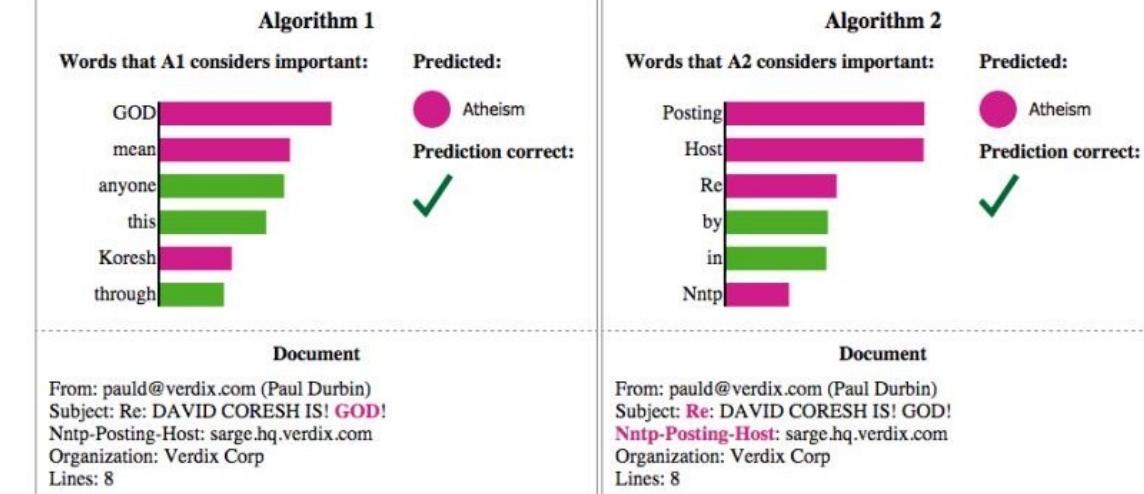
Hendrik Strobelt, Sebastian Gehrmann, Michael Behrisch, Adam Perer, Hanspeter Pfister, Alexander M. Rush: Seq2seq-Vis: A Visual Debugging Tool for Sequence-to-Sequence Models. IEEE Trans. Vis. Comput. Graph. 25(1): 353-363 (2019)

- Fine-grained explanations are in the form of:
- texts in a real-world dataset;
 - Numerical scores

Example #3 of 6

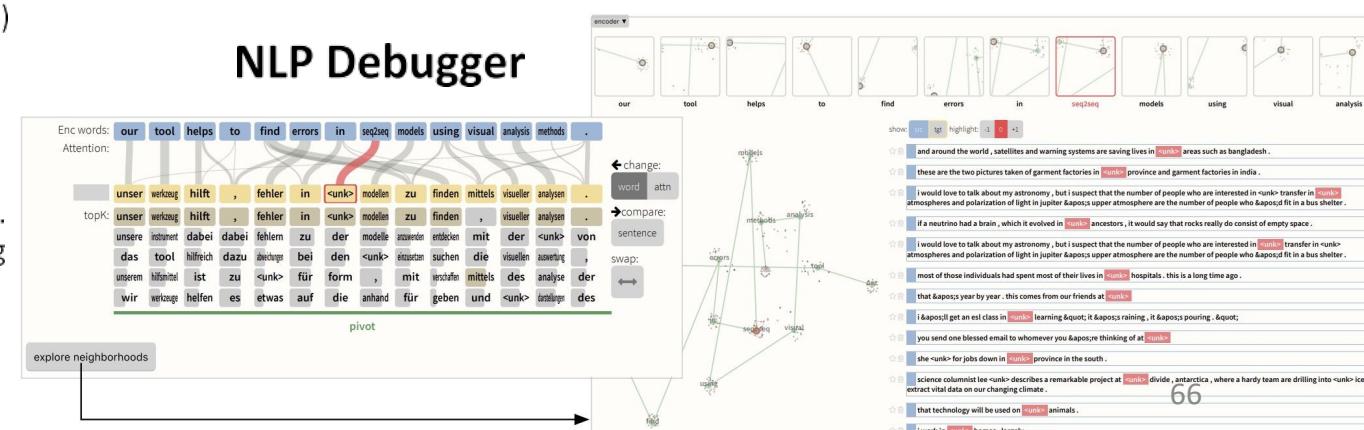
True Class: Atheism

Instructions Previous Next



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

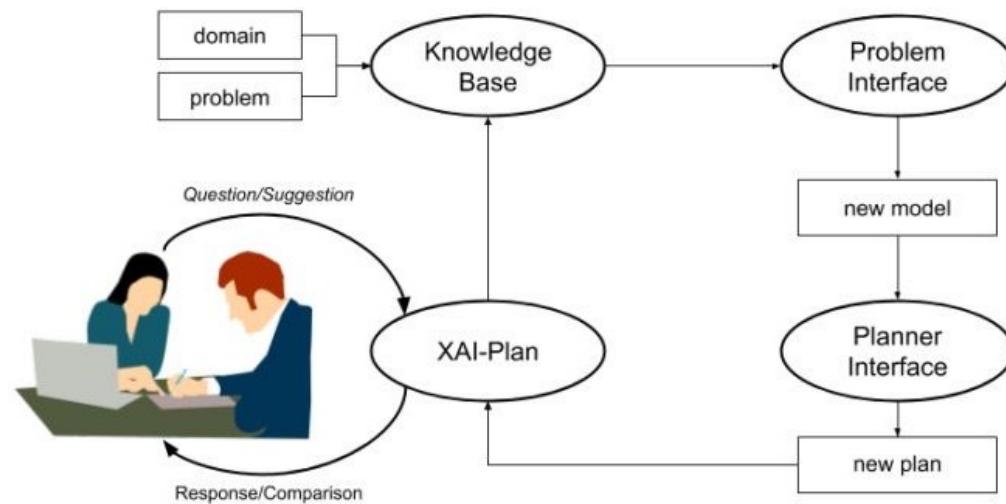


Overview of Explanation in Different AI Fields (6)

• Planning and Scheduling

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

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



XAI Plan

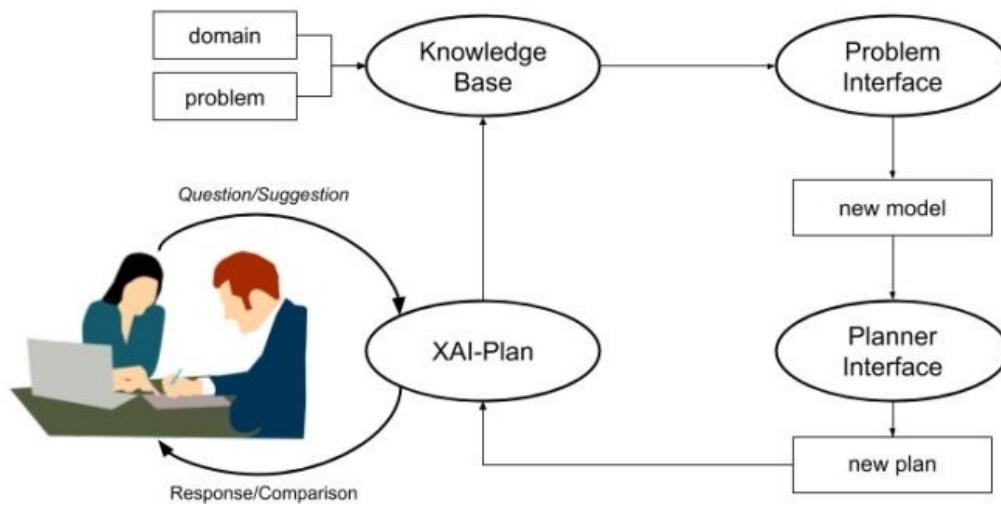
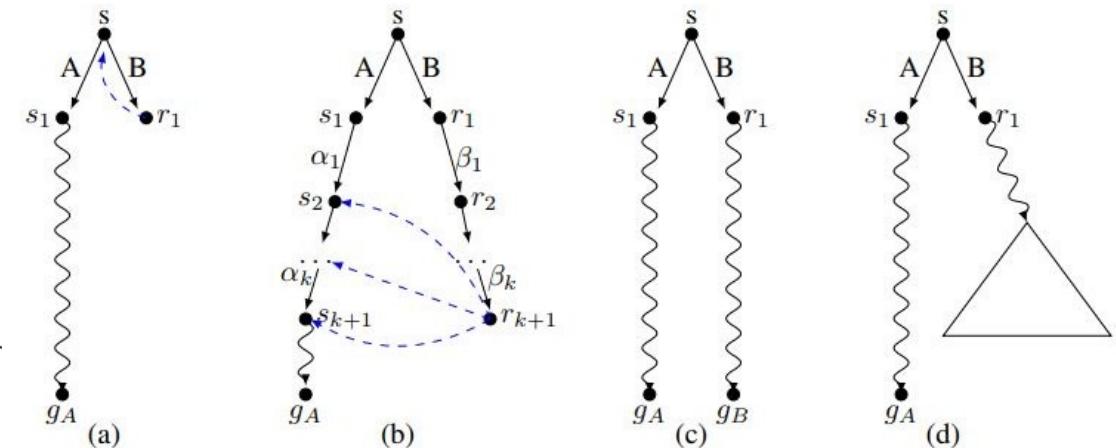
Rita Borgo, Michael Cashmore, Daniele Magazzeni: Towards Providing Explanations for AI Planner Decisions. CoRR abs/1810.06338 (2018)

Overview of Explanation in Different AI Fields (6)

• Planning and Scheduling

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

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



XAI Plan

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

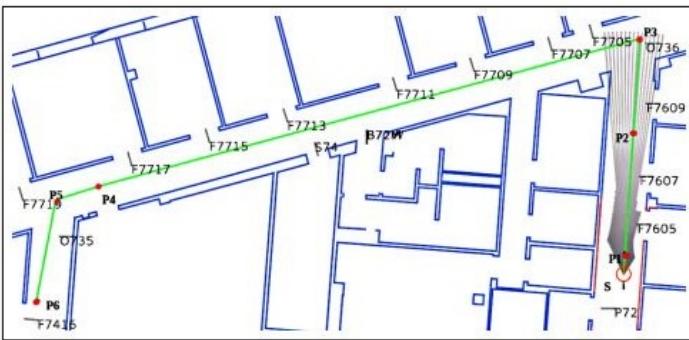
Human-in-the-loop Planning

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

(Manual) Plan Comparison

Overview of Explanation in Different AI Fields (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

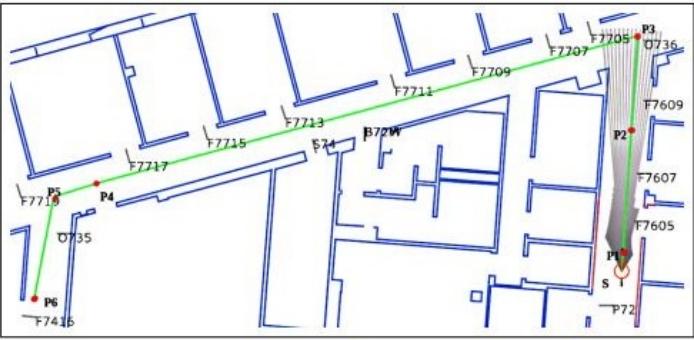
Narration of Autonomous Robot Experience

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



		Abstraction, A				
Specificity, S	Level 1		Level 2		Level 3	
	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?

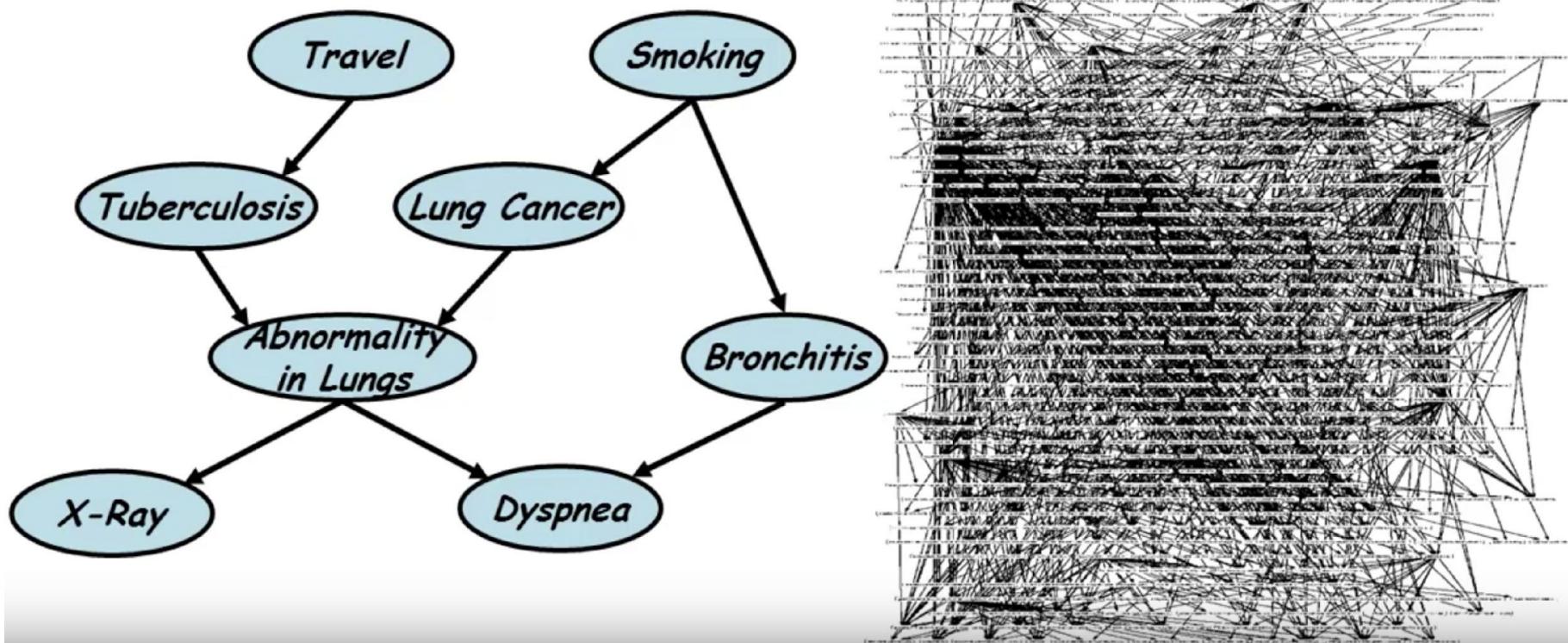
Robot: Here is the histogram of all training examples that
reached this leaf. 80% of examples where this area was
above 20 cm predicted the appropriate action to be “drive
forward”.

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



Explainable Machine Learning (from a Machine Learning Perspective)

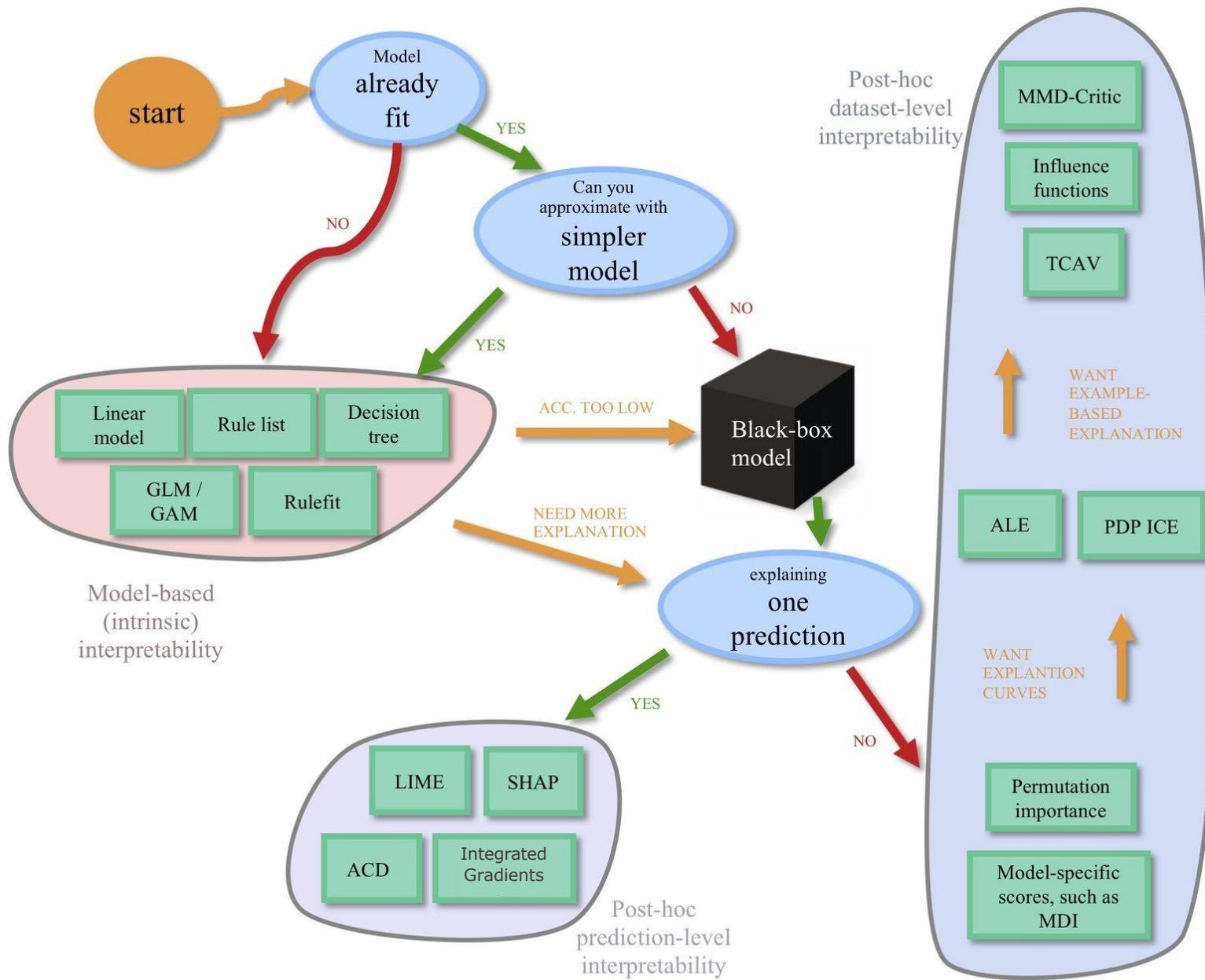
Achieving Explainable AI

Approach 1: Post-hoc explain a given AI model

- Individual prediction explanations in terms of input features, influential examples, concepts, local decision rules
- Global prediction explanations in terms of entire model in terms of partial dependence plots, global feature importance, global decision rules

Approach 2: Build an interpretable model

- Logistic regression, Decision trees, Decision lists and sets, Generalized Additive Models (GAMs)



interpretability cheat-sheet

[View on github](#)

Based on [this interpretability review](#) and the [sklearn cheat-sheet](#). More in [this book](#) + these [slides](#).

Summaries and links to code

[RuleFit](#) – automatically add features extracted from a small tree to a linear model

[LIME](#) – linearly approximate a model at a point

[SHAP](#) – find relative contributions of features to a prediction

[ACD](#) – hierarchical feature importances for a DNN prediction

[Text](#) – DNN generates text to explain a DNN's prediction (sometimes not faithful)

[Permutation importance](#) – permute a feature and see how it affects the model

[ALE](#) – perturb feature value of nearby points and see how outputs change

[PDP ICE](#) – vary feature value of all points and see how outputs change

[TCAV](#) – see if representations of certain points learned by DNNs are linearly separable

[Influence functions](#) – find points which highly influence a learned model

[MMD-CRITIC](#) – find a few points which summarize classes

Achieving Explainable AI

Approach 1: Post-hoc explain a given AI model

- **Individual prediction explanations** in terms of **input features, influential examples, concepts, local decision rules**
- **Global prediction explanations** in terms of entire model in terms of **partial dependence plots, global feature importance, global decision rules**

Approach 2: Build an interpretable model

- Logistic regression, Decision trees, Decision lists and sets, Generalized Additive Models (GAMs)



Top label: "**clog**"

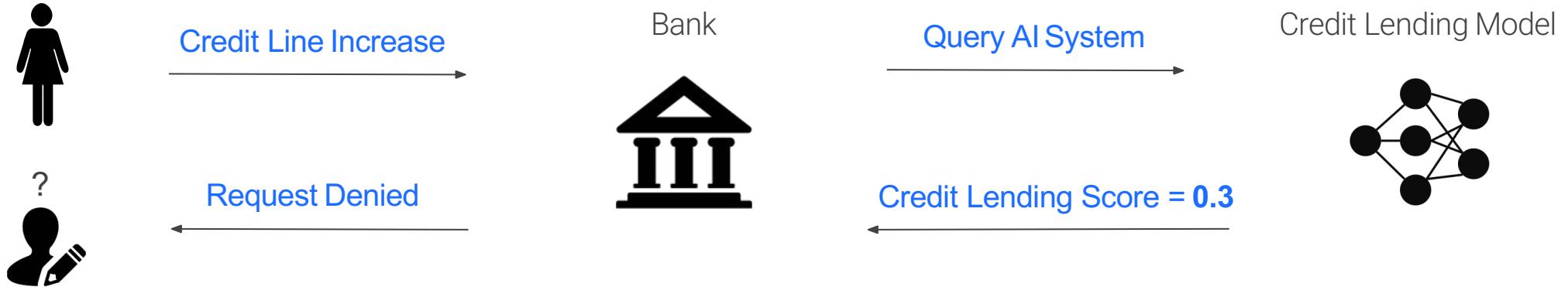
Why did the network label this image as "**clog**"?



Top label: "**fireboat**"

Why did the network label this image as "**fireboat**"?

Credit Lending in a black-box ML world



Why? Why not?

How?

Fair lending laws [ECOA, FCRA] require credit decisions to be explainable

The Attribution Problem

Attribute a model's prediction on an input to features of the input

Examples:

- Attribute an object recognition network's prediction to its pixels
- Attribute a text sentiment network's prediction to individual words
- Attribute a lending model's prediction to its features

A reductive formulation of “why this prediction” **but surprisingly useful**

Application of Attributions

- **Debugging model predictions**
E.g., Attribution an image misclassification to the pixels responsible for it
- **Generating an explanation for the end-user**
E.g., Expose attributions for a lending prediction to the end-user
- **Analyzing model robustness**
E.g., Craft adversarial examples using weaknesses surfaced by attributions
- **Extract rules from the model**
E.g., Combine attribution to craft rules (pharmacophores) capturing prediction logic of a drug screening network

Next few slides

We will cover the following **attribution methods****

- Ablations
- Gradient based methods (specific to differentiable models)
- Score Backpropagation based methods (specific to NNs)

We will also discuss game theory (Shapley value) in attributions

**Not a complete list!

See Ancona et al. [ICML 2019], Guidotti et al. [arxiv 2018] for a comprehensive survey

Ablations

Drop each feature and attribute the change in prediction to that feature

Pros:

- Simple and intuitive to interpret

Cons:

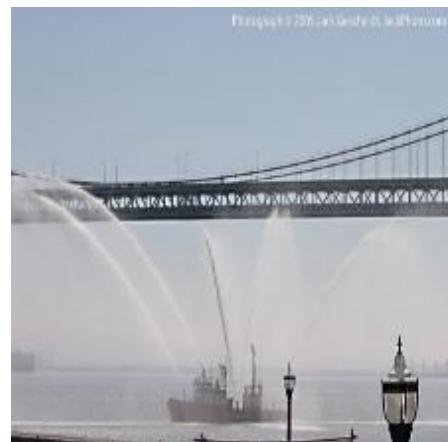
- Unrealistic inputs
- Improper accounting of interactive features
- Can be computationally expensive



Feature*Gradient

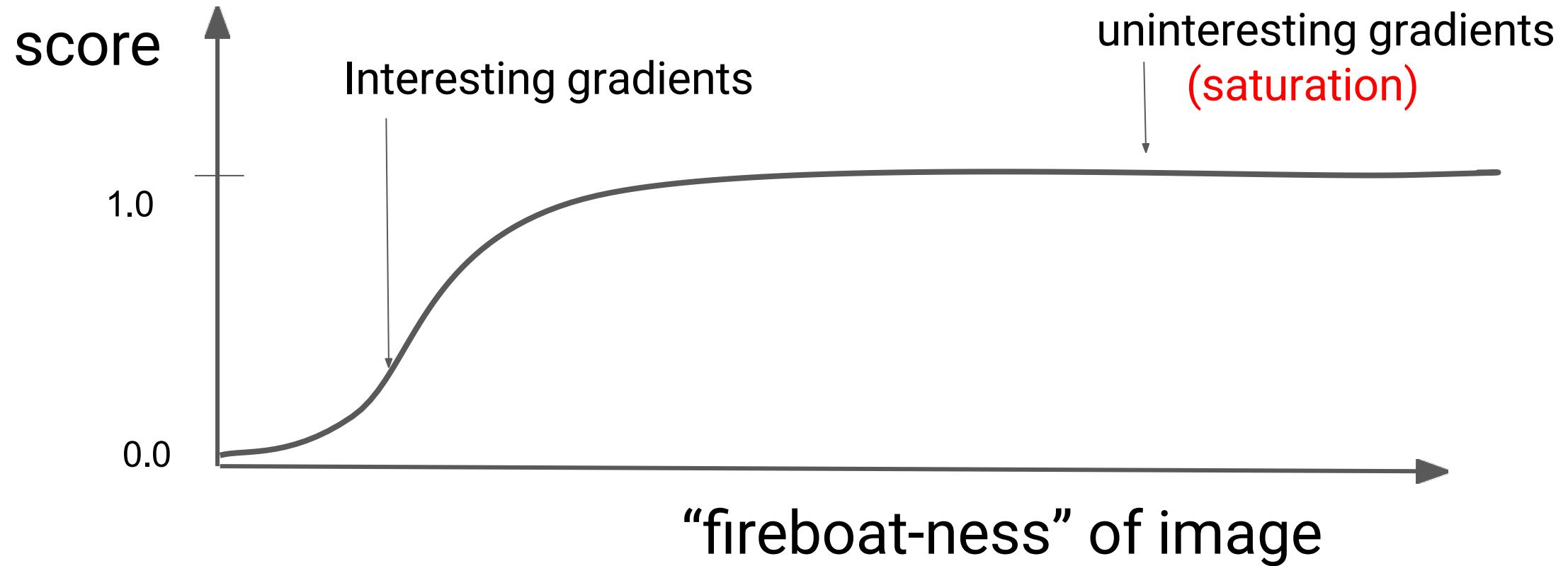
Attribution to a feature is feature value times gradient, i.e., $x_i^* \partial y / \partial x_i$

- Gradient captures sensitivity of output w.r.t. feature
- Equivalent to Feature*Coefficient for linear models
 - First-order Taylor approximation of non-linear models
- Popularized by SaliencyMaps [NIPS 2013], Baehrens et al. [JMLR 2010]



Gradients in the vicinity of the input seem like noise?

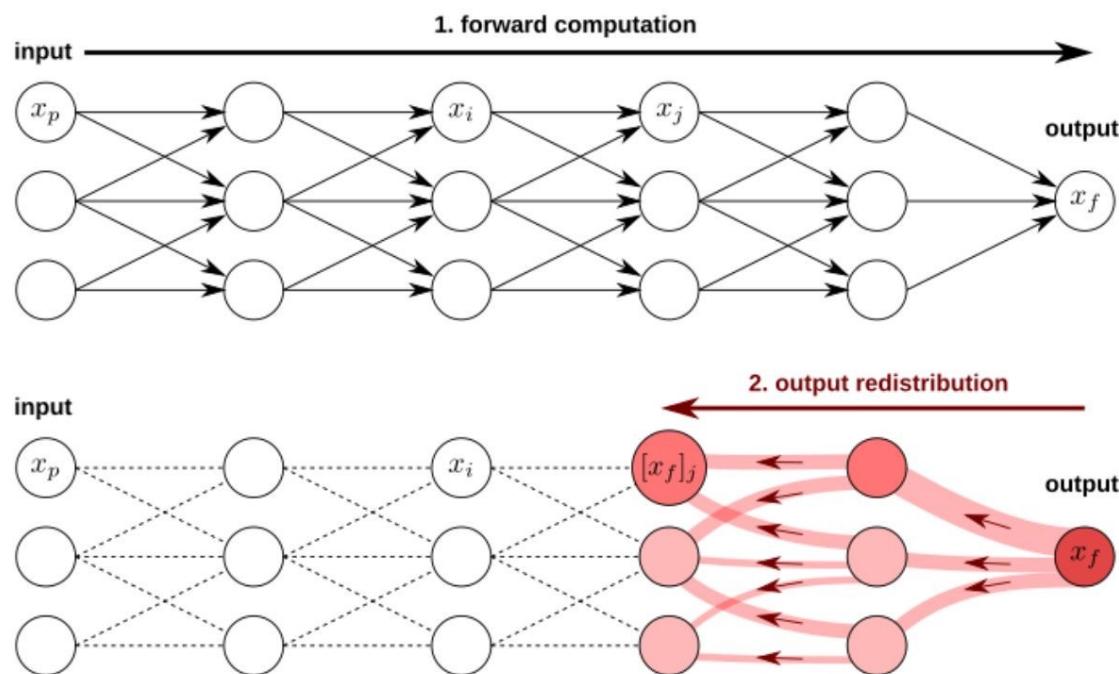
Local linear approximations can be too local



Score Back-Propagation based Methods

Re-distribute the prediction score through the neurons in the network

- LRP [JMLR 2017], DeepLift [ICML 2017], Guided BackProp [ICLR 2014]



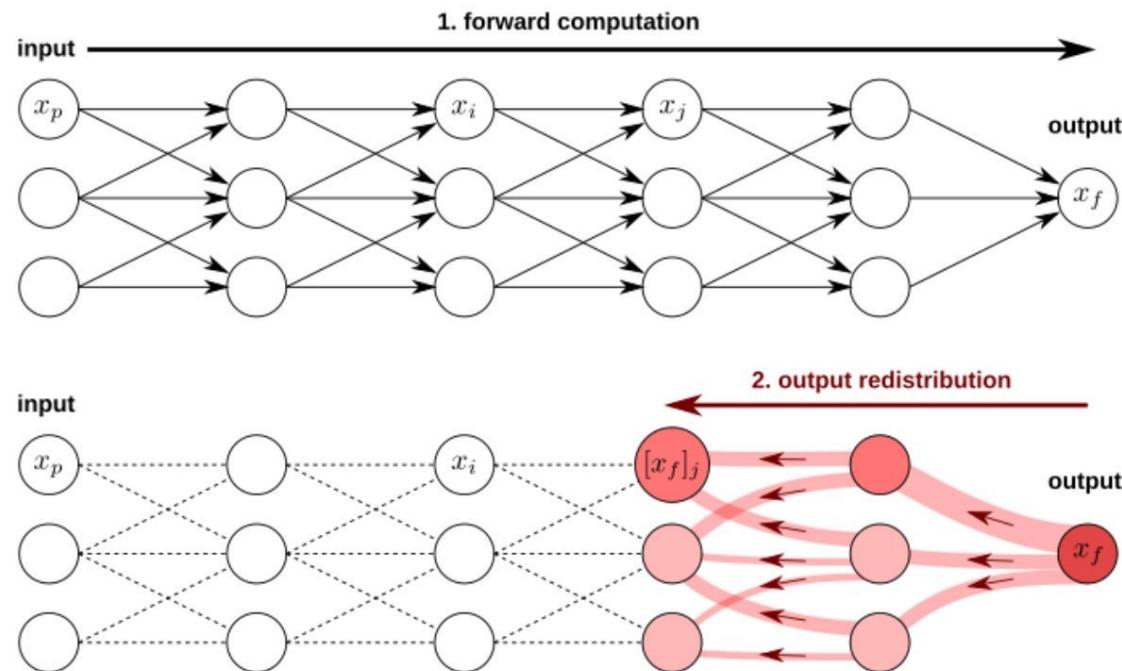
Easy case: Output of a neuron is a linear function of previous neurons (i.e., $n_i = \sum w_{ij} * n_j$)
e.g., the logit neuron

- Re-distribute the contribution in proportion to the coefficients w_{ij}

Score Back-Propagation based Methods

Re-distribute the prediction score through the neurons in the network

- LRP [JMLR 2017], DeepLift [ICML 2017], Guided BackProp [ICLR 2014]



Tricky case: Output of a neuron is a **non-linear** function, e.g., ReLU, Sigmoid, etc.

- **Guided BackProp:** Only consider ReLUs that are on (linear regime), and which contribute positively
- **LRP:** Use first-order Taylor decomposition to linearize activation function
- **DeepLift:** Distribute activation difference relative a reference point in proportion to edge weights

Score Back-Propagation based Methods

Re-distribute the prediction score through the neurons in the network

- LRP [JMLR 2017], DeepLift [ICML 2017], Guided BackProp [ICLR 2014]

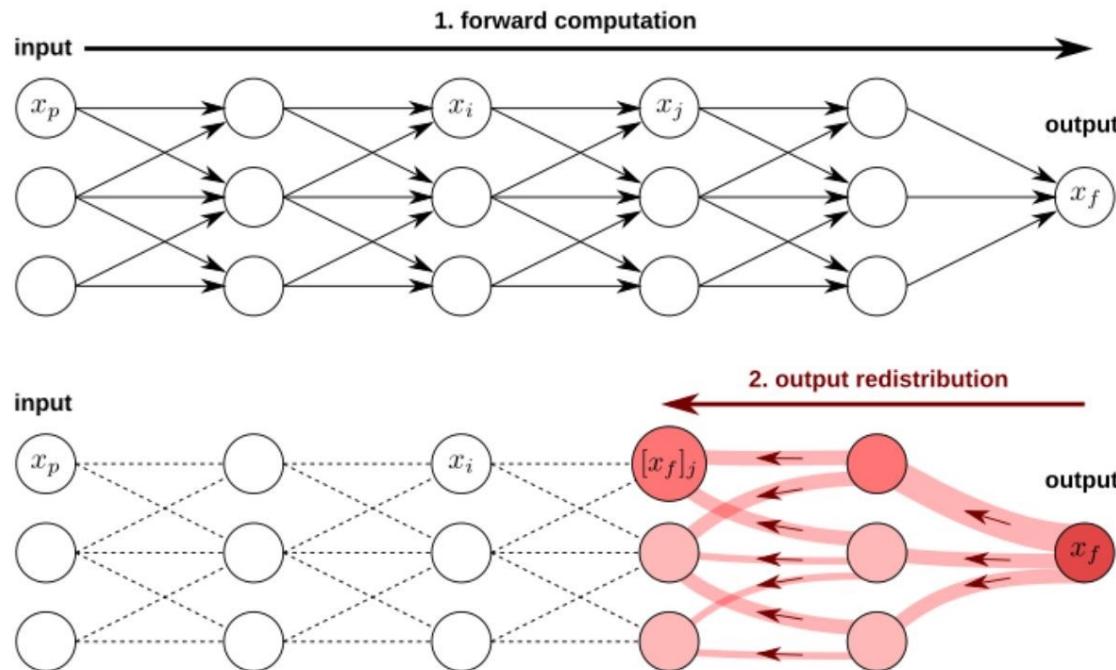


Image credit heatmapping.org

Pros:

- Conceptually simple
- Methods have been empirically validated to yield sensible result

Cons:

- Hard to implement, requires instrumenting the model
- **Often breaks implementation invariance**

Think: $F(x, y, z) = x * y * z$ and
 $G(x, y, z) = x * (y * z)$

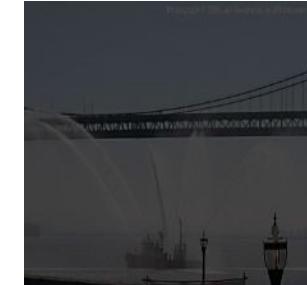
Baselines and additivity

- When we decompose the score via backpropagation, we imply a normative alternative called a **baseline**
 - “Why $\text{Pr}(\text{fireboat}) = 0.91$ [instead of 0.00]”
- Common choice is an **informationless input for the model**
 - E.g., Black image for image models
 - E.g., Empty text or zero embedding vector for text models
- **Additive** attributions explain $F(\text{input}) - F(\text{baseline})$ in terms of input features

Another approach: gradients at many points



Baseline



... scaled inputs ...

Input



... gradients of scaled inputs

Integrated Gradients [ICML 2017]

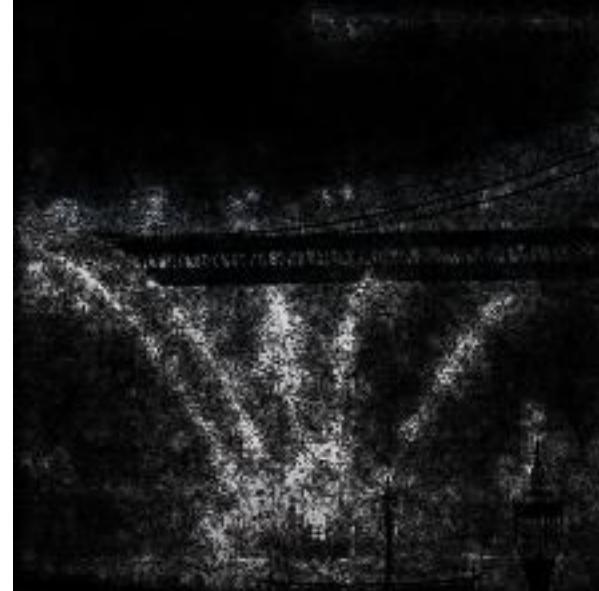
Integrate the gradients along a **straight-line path from baseline to input**

$$IG(\text{input}, \text{base}) ::= (\text{input} - \text{base}) * \int_{0-1} \nabla F(\alpha * \text{input} + (1-\alpha) * \text{base}) d\alpha$$

Original image



Integrated Gradients





Integrated Gradients in action

Why is this image labeled as “clog”?

Original image



“Clog”



Why is this image labeled as “clog”?

Original image



Integrated Gradients
(for label “clog”)

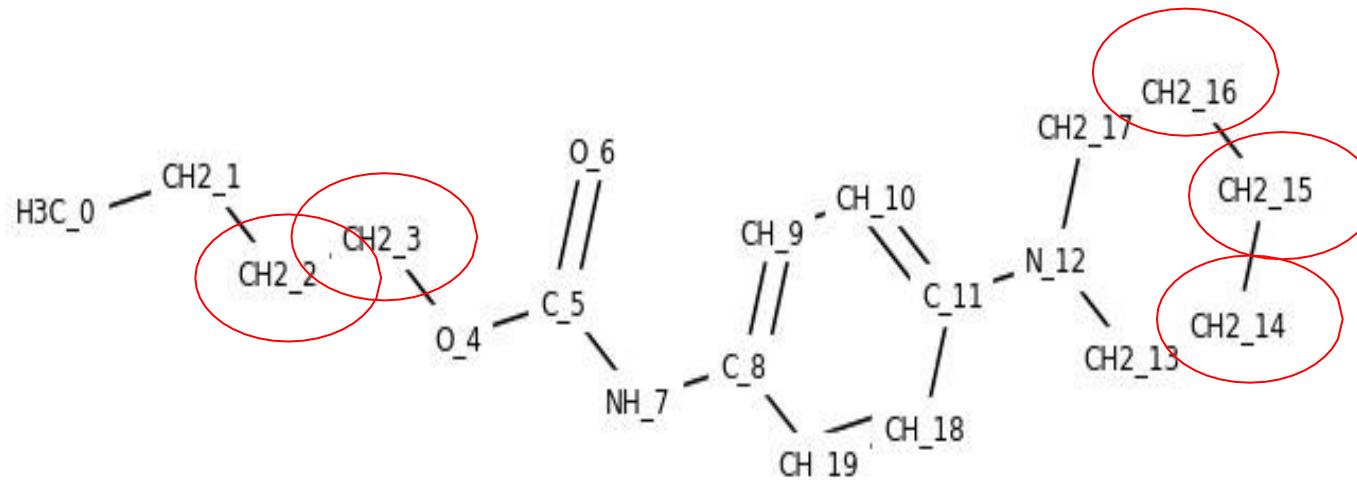


“Clog”



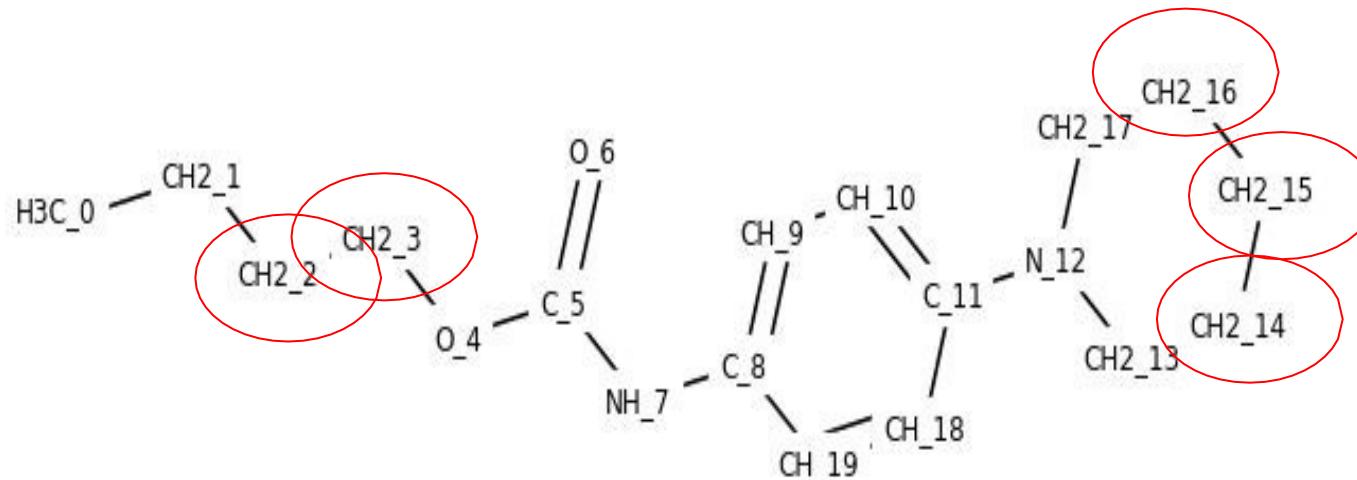
Detecting an architecture bug

- Deep network [Kearns, 2016] predicts if a molecule binds to certain DNA site
- **Finding:** Some atoms had identical attributions despite different connectivity



Detecting an architecture bug

- Deep network [Kearns, 2016] predicts if a molecule binds to certain DNA site
- **Finding:** Some atoms had identical attributions despite different connectivity

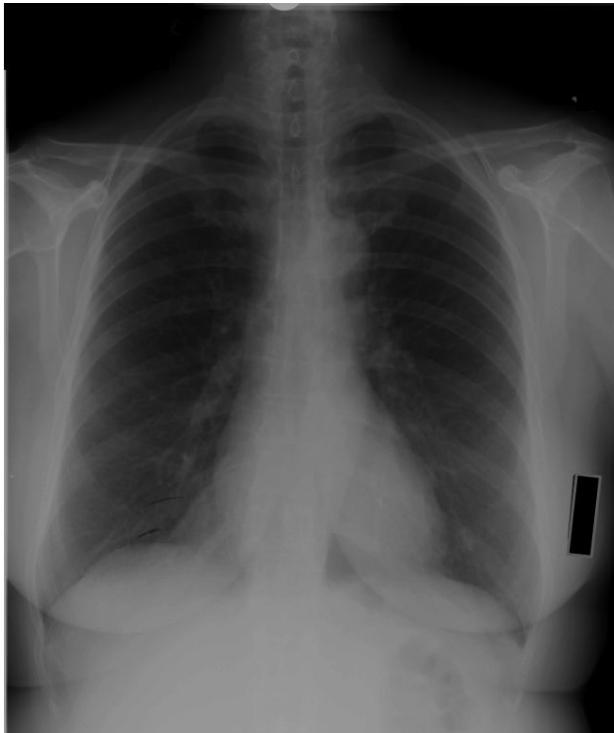


- **Bug:** The architecture had a bug due to which the convolved bond features did not affect the prediction!

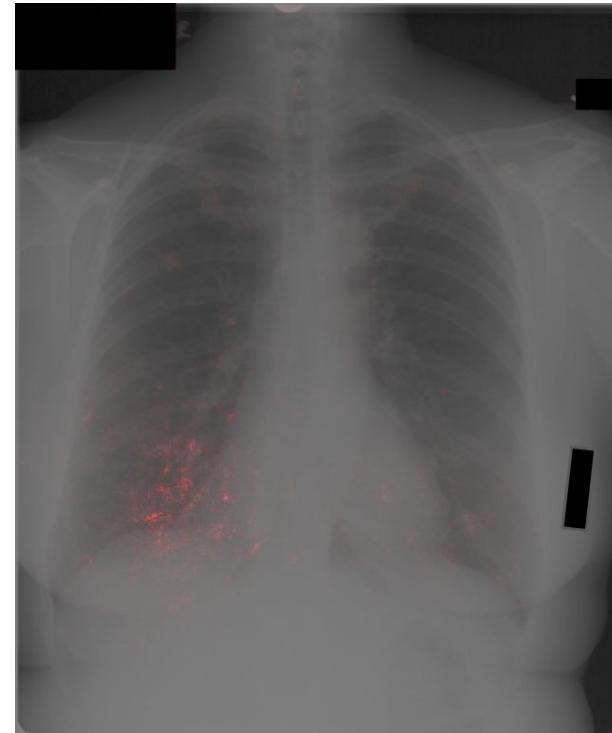
Detecting a data issue

- Deep network predicts various diseases from chest x-rays

Original image

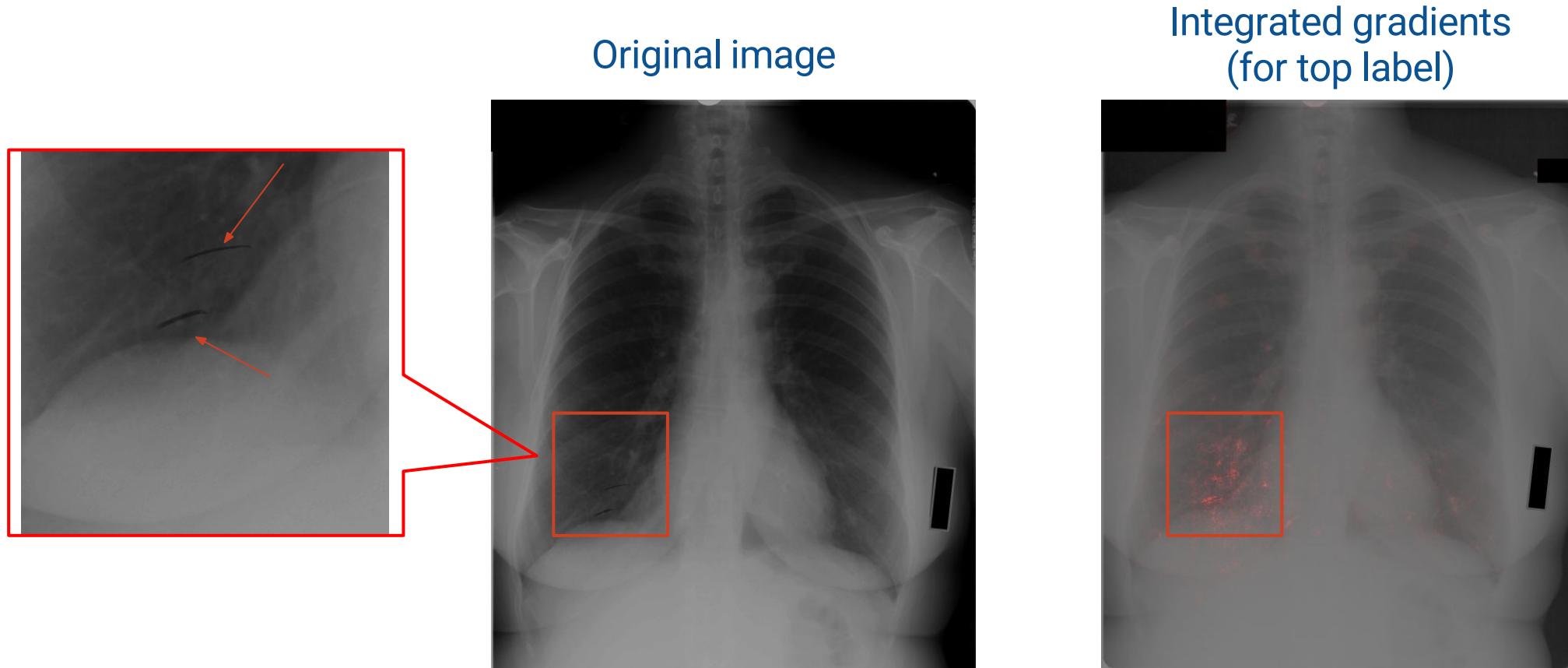


Integrated gradients
(for top label)



Detecting a data issue

- Deep network predicts various diseases from chest x-rays
- **Finding:** Attributions fell on radiologist's markings (rather than the pathology)





Cooperative game theory in attributions

Shapley Value [Annals of Mathematical studies, 1953]

Classic result in game theory on distributing gain in a **coalition game**

- **Coalition Games**

- Players collaborating to generate some **gain** (think: revenue)
- Set function $v(S)$ determining the gain for any subset S of players

Shapley Value [Annals of Mathematical studies, 1953]

Classic result in game theory on distributing gain in a **coalition game**

- **Coalition Games**
 - Players collaborating to generate some **gain** (think: revenue)
 - Set function $v(S)$ determining the gain for any subset S of players
- **Shapley Values** are a fair way to attribute the total gain to the players based on their contributions
 - Concept: **Marginal contribution** of a player to a subset of other players ($v(S \cup \{i\}) - v(S)$)
 - Shapley value for a player is a **specific weighted aggregation** of its marginal over all possible subsets of other players

$$\text{Shapley Value for player } i = \sum_{S \subseteq N} w(S) * (v(S \cup \{i\}) - v(S))$$

$$(\text{where } w(S) = N! / |S|! (N - |S| - 1)!)$$

Shapley Value Justification

Shapley values are unique under four simple axioms

- **Dummy:** If a player never contributes to the game then it must receive zero attribution
- **Efficiency:** Attributions must add to the total gain
- **Symmetry:** Symmetric players must receive equal attribution
- **Linearity:** Attribution for the (weighted) sum of two games must be the same as the (weighted) sum of the attributions for each of the games

Shapley Values for Explaining ML models

SHAP [NeurIPS 2018], QII [S&P 2016], Strumbelj & Konenko [JMLR 2009]

- Define a coalition game for each model input X
 - **Players are the features in the input**
 - **Gain is the model prediction (output), i.e., gain = $F(X)$**
- Feature attributions are the Shapley values of this game

Shapley Values for Explaining ML models

SHAP [NeurIPS 2018], QII [S&P 2016], Strumbelj & Konenko [JMLR 2009]

- Define a coalition game for each model input X
 - **Players are the features in the input**
 - **Gain is the model prediction (output), i.e., gain = $F(X)$**
- Feature attributions are the Shapley values of this game

Challenge: Shapley values require the gain to be defined for all subsets of players

- What is the prediction when **some players (features) are absent?**
i.e., what is $F(x_1, \langle \text{absent} \rangle, x_3, \dots, \langle \text{absent} \rangle)$?

Modeling Feature Absence

Key Idea: Take the expected prediction when the (absent) feature is sampled from a certain distribution.

Different approaches choose different distributions

- [SHAP, NIPS 2018] Use conditional distribution w.r.t. the present features
- [QII, S&P 2016] Use marginal distribution
- [Strumbelj et al., JMLR 2009] Use uniform distribution

Computing Shapley Values

Exact Shapley value computation is **exponential in the number of features**

- Shapley values can be expressed as an expectation of marginals

$$\phi(i) = E_{S \sim D} [\text{marginal}(S, i)]$$

- Sampling-based methods can be used to approximate the expectation
- See: “[Computational Aspects of Cooperative Game Theory](#)”, Chalkiadakis et al. 2011
- The method is still computationally infeasible for models with hundreds of features, e.g., image models

Non-atomic Games: Aumann-Shapley Values and IG

- *Values of Non-Atomic Games* (1974): Aumann and Shapley extend their method → players can contribute fractionally
- Aumann-Shapley values calculated by integrating along a straight-line path...
same as Integrated Gradients!
- IG through a game theory lens: continuous game, feature absence is modeled by replacement with a baseline value
- Axiomatically justified as a result:
 - Integrated Gradients is the unique path-integral method satisfying: **Sensitivity**, **Insensitivity**, **Linearity preservation**, **Implementation invariance**, **Completeness**, and **Symmetry**

Lessons learned: baselines are important

Baselines (or Norms) are essential to explanations [\[Kahneman-Miller 86\]](#)

- E.g., A man suffers from indigestion. Doctor blames it to a stomach ulcer. Wife blames it on eating turnips. Both are correct relative to their baselines.
- The baseline may also be an important analysis knob.

Attributions are **contrastive**, whether we think about it or not.



Some limitations
and caveats for
attributions

Attributions don't explain everything

Some things that are missing:

- Feature interactions (ignored or averaged out)
- What training examples influenced the prediction (training agnostic)
- Global properties of the model (prediction-specific)

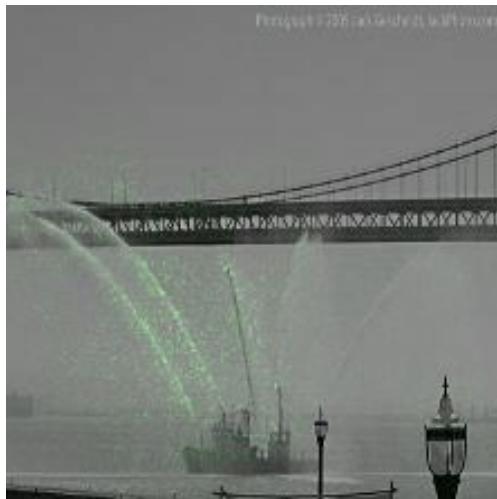
An instance where attributions are useless:

- A model that predicts TRUE when there are **even number** of black pixels and FALSE otherwise

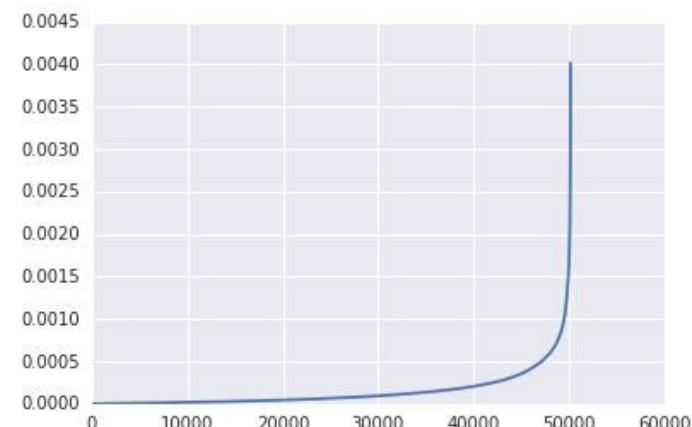
Attributions are for human consumption

- **Humans** interpret attributions and generate insights
 - Doctor maps attributions for x-rays to pathologies
- **Visualization** matters as much as the attribution technique

Naive scaling of attributions from 0 to 255



Attributions have a **large range** and **long tail** across pixels



After clipping attributions at 99% to reduce range





Other individual
prediction
explanation
methods

Local Interpretable Model-agnostic Explanations

(Ribeiro et al. KDD 2016)

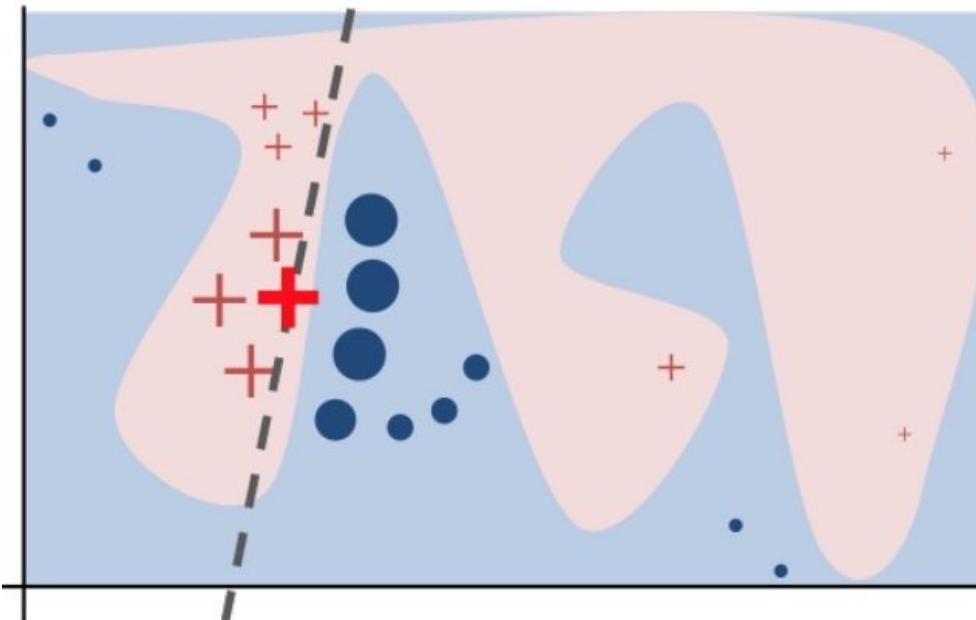
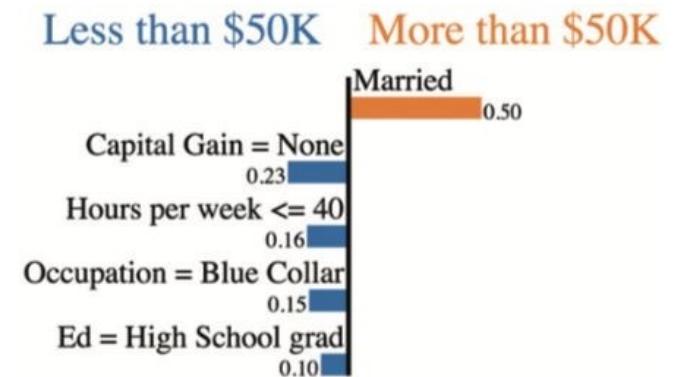


Figure credit: Ribeiro et al. KDD 2016

28 < Age \leq 37
Workclass = Private
Education = High School grad
Marital Status = Married
Occupation = Blue-Collar
Relationship = Husband
Race = White
Sex = Male
Capital Gain = None
Capital Loss = Low
Hours per week \leq 40.00
Country = United-States

$$P(\text{Salary} > \$50K) = 0.57$$

(a) Instance and prediction



(b) LIME explanation

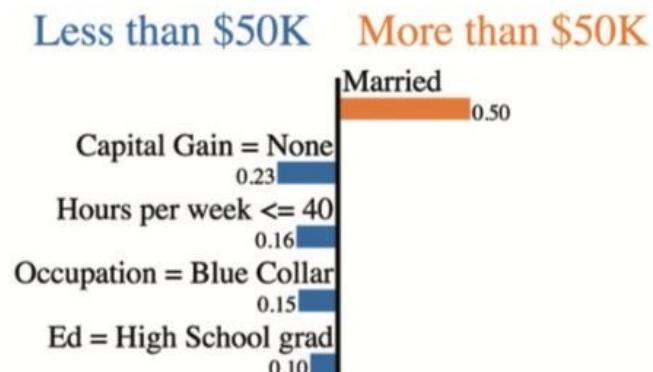
Figure credit: Anchors: High-Precision Model-Agnostic Explanations. Ribeiro et al. AAAI 2018

Anchors

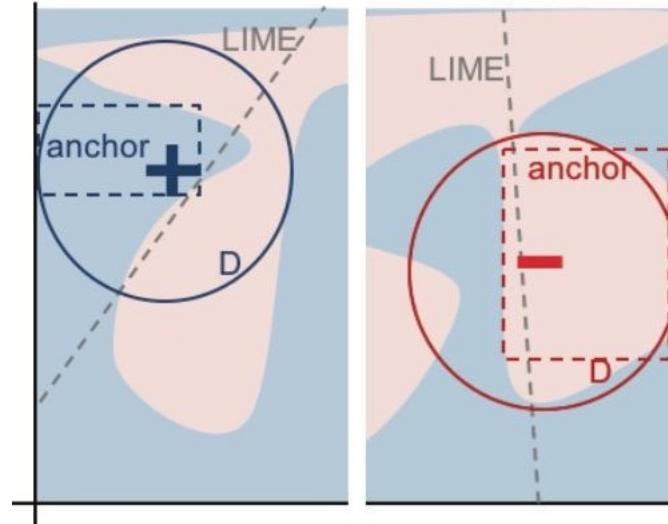
28 < Age ≤ 37
Workclass = Private
Education = High School grad
Marital Status = Married
Occupation = Blue-Collar
Relationship = Husband
Race = White
Sex = Male
Capital Gain = None
Capital Loss = Low
Hours per week ≤ 40.00
Country = United-States

$P(\text{Salary} > \$50K) = 0.57$

(a) Instance and prediction



(b) LIME explanation



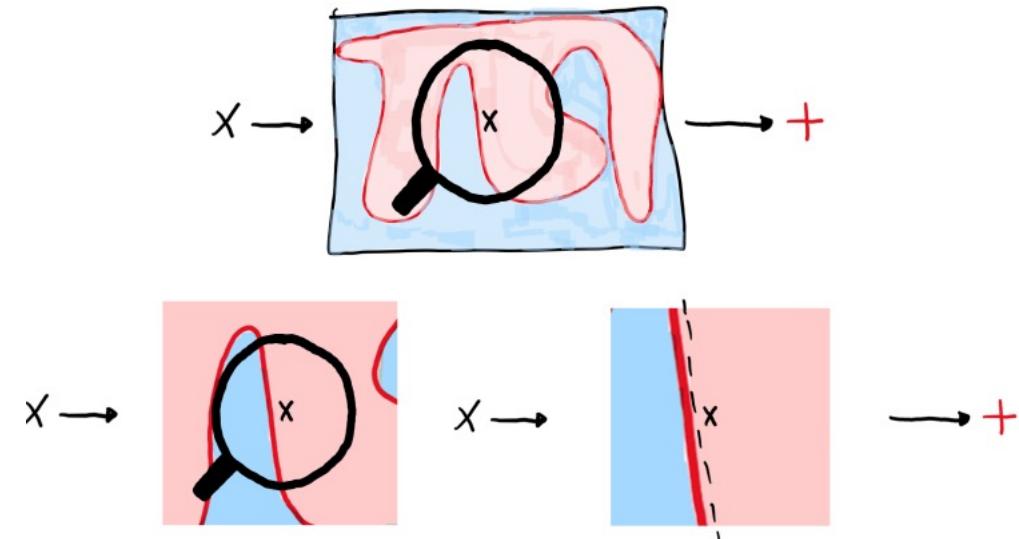
**IF Country = United-States AND Capital Loss = Low
AND Race = White AND Relationship = Husband
AND Married AND 28 < Age ≤ 37
AND Sex = Male AND High School grad
AND Occupation = Blue-Collar
THEN PREDICT Salary > \$50K**

(c) An *anchor* explanation

Figure credit: Anchors: High-Precision Model-Agnostic Explanations. Ribeiro et al. AAAI 2018

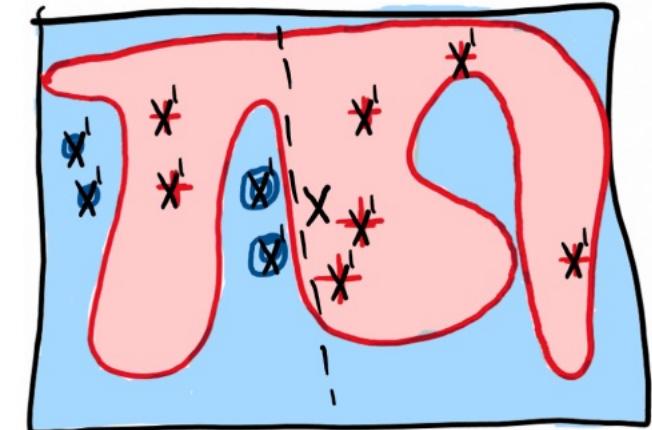
Local Interpretable Model-Agnostic Explanations (LIME)

- Global explanations can be too complex
- Zoom in to examine local interpretability
- Summary:
 - Simplify a global model by perturbing input to see how predictions change
 - Approximate underlying model learned on these perturbations



Local Interpretable Model-Agnostic Explanations (LIME)

- Steps:
 - Sample points around X
 - Get predictions from our original model (complex)
 - Weight samples according to our distance from x (cos for text, L2 for images)
 - Learn a simple model from our weighted samples
 - Utilize simple model for better interpretability!



Influence functions

- Trace a model's prediction through the learning algorithm and back to its training data
- Training points “responsible” for a given prediction

Test image

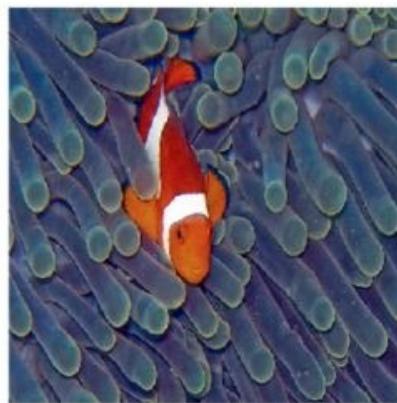
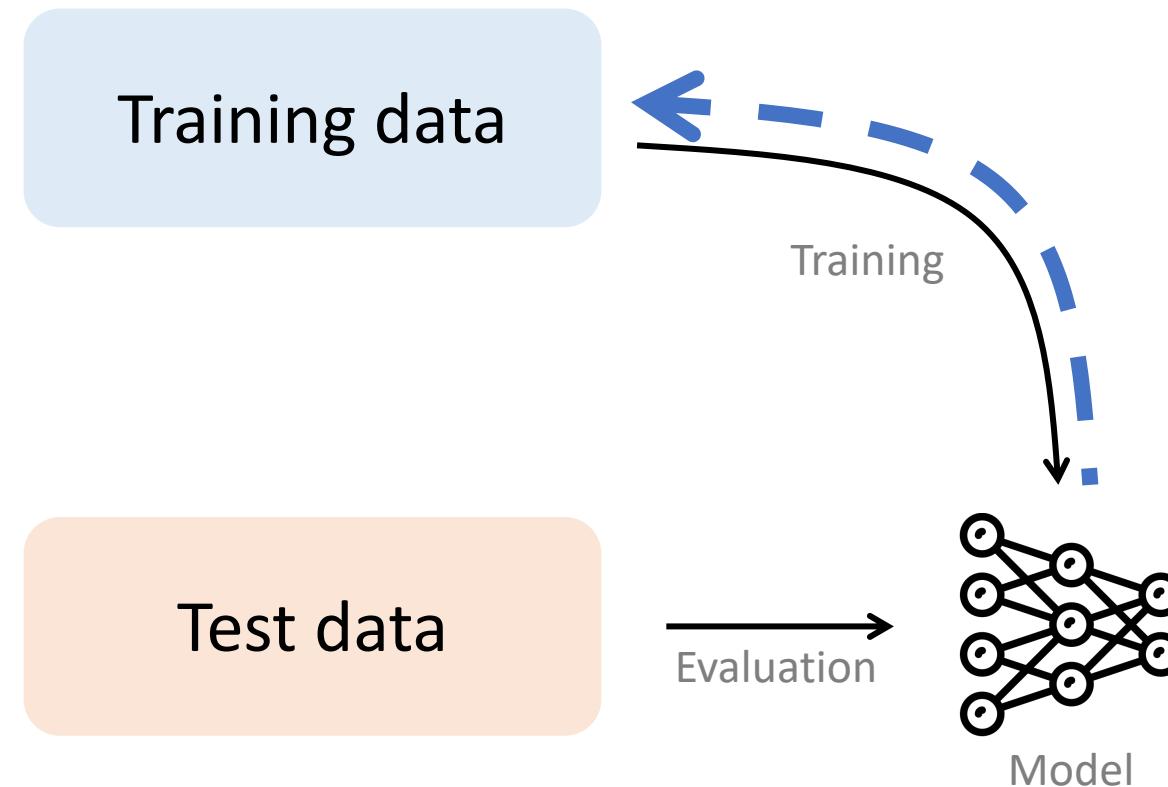
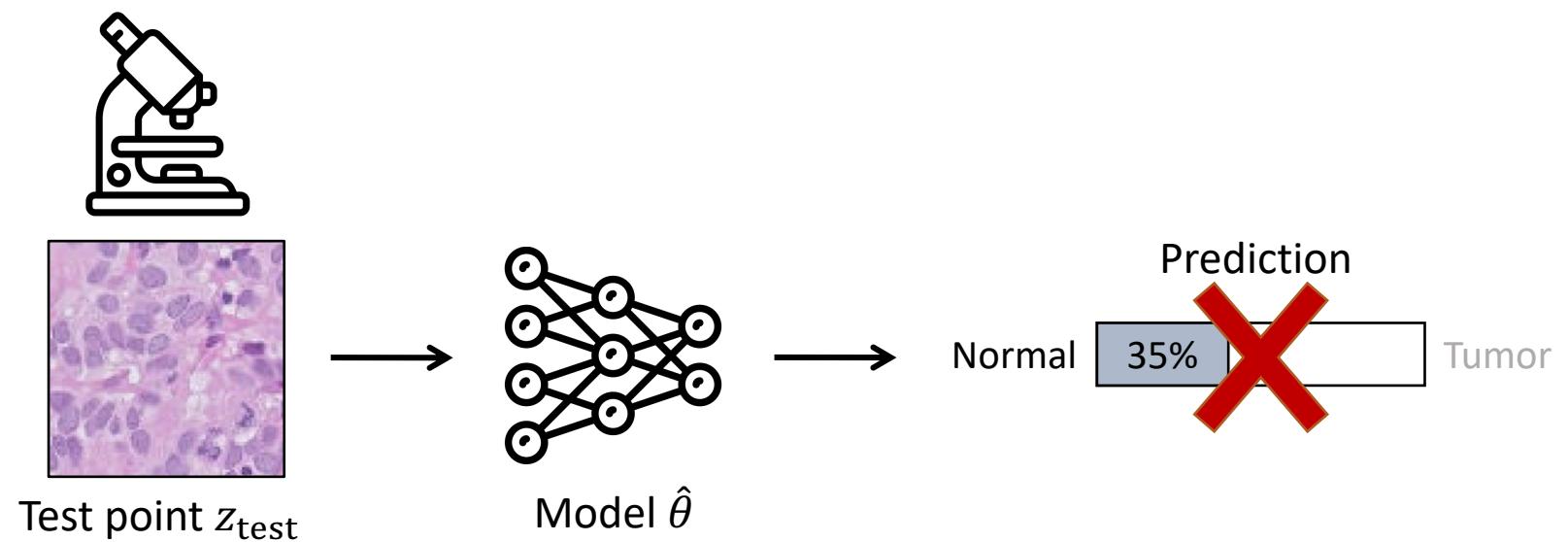


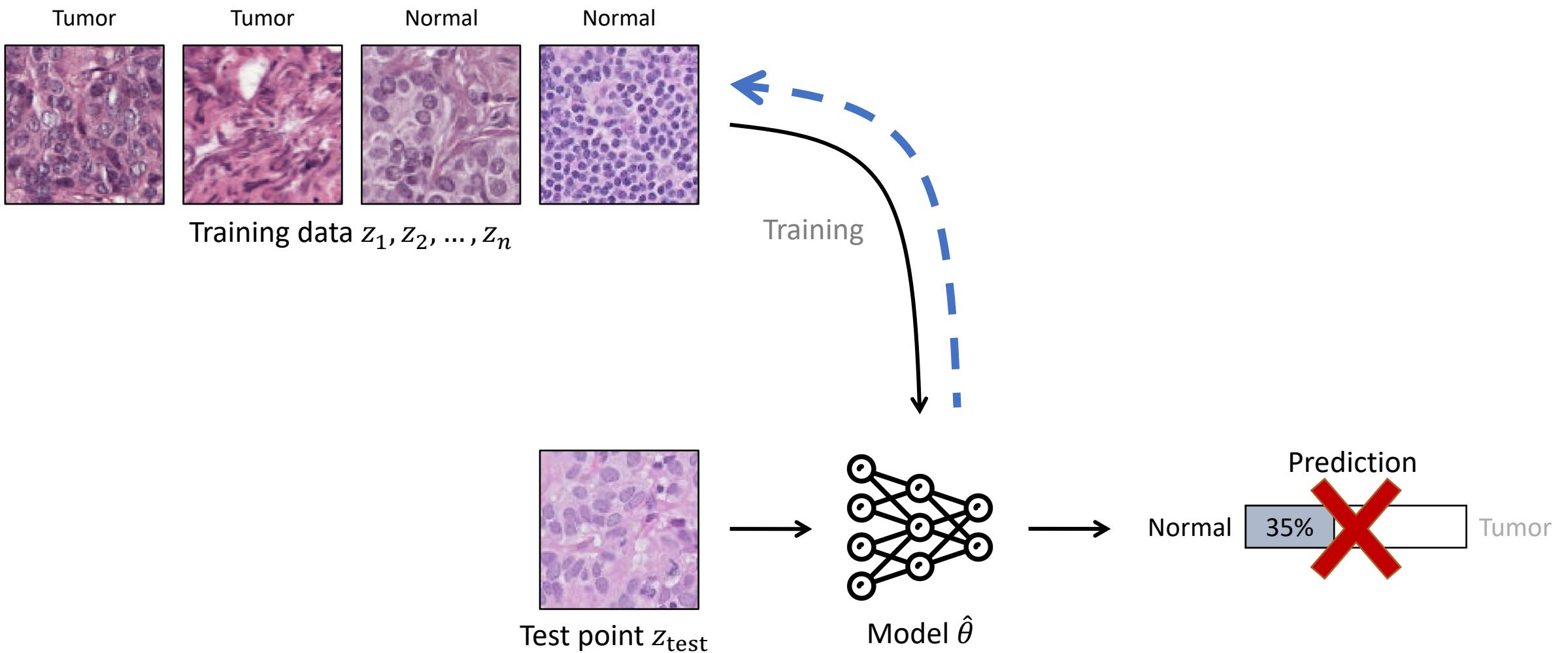
Figure credit: Understanding Black-box Predictions via Influence Functions. Koh and Liang. ICML 2017

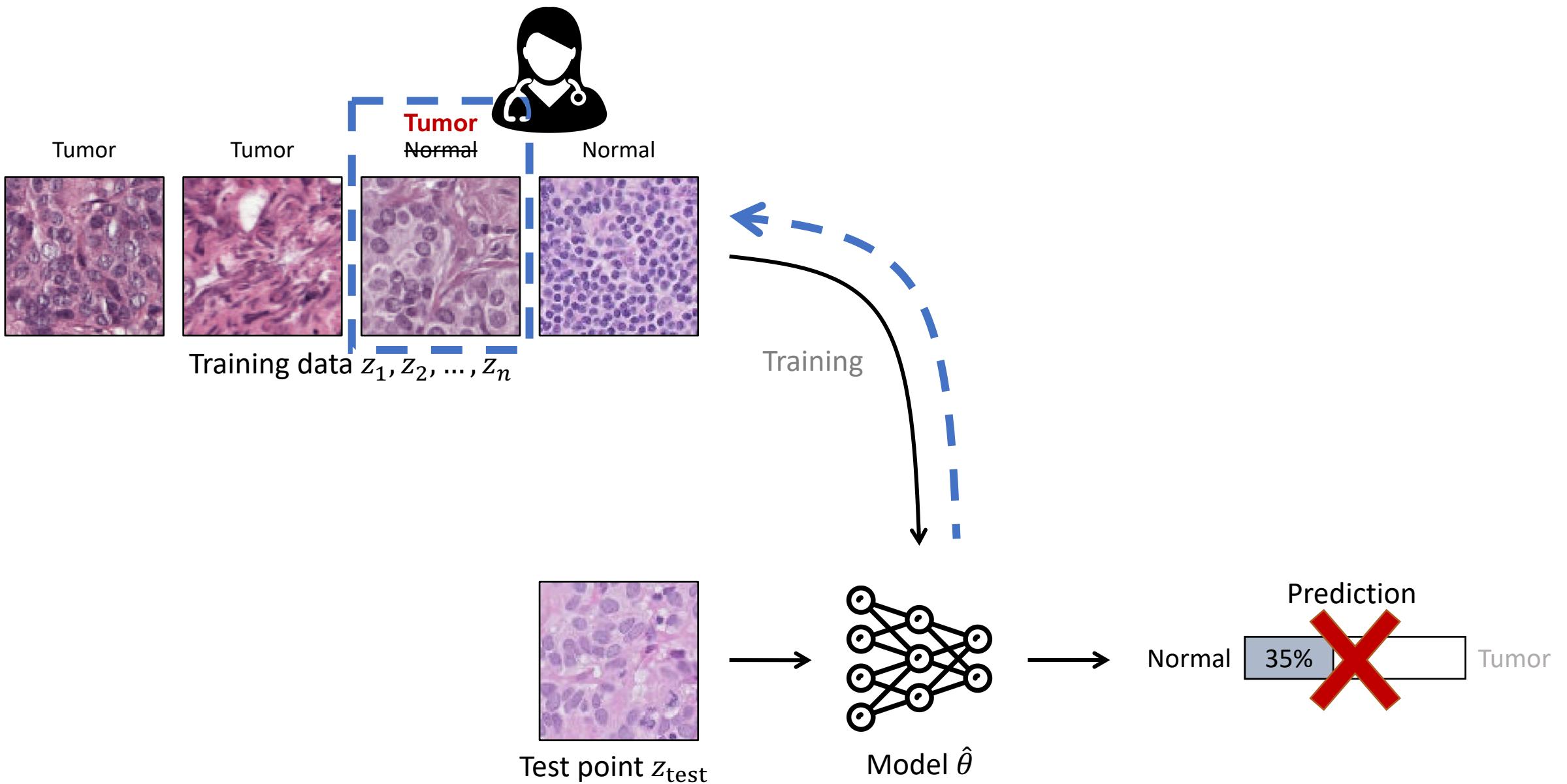
Influence functions: Link model to training data

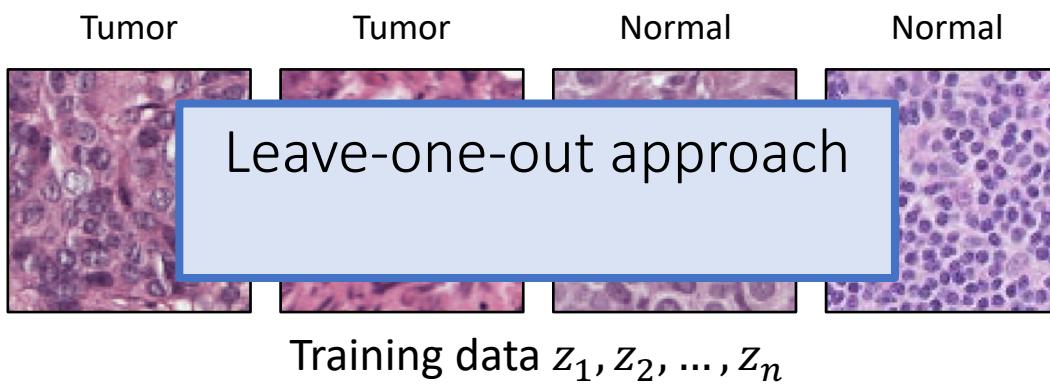


Example: Dataset debugging

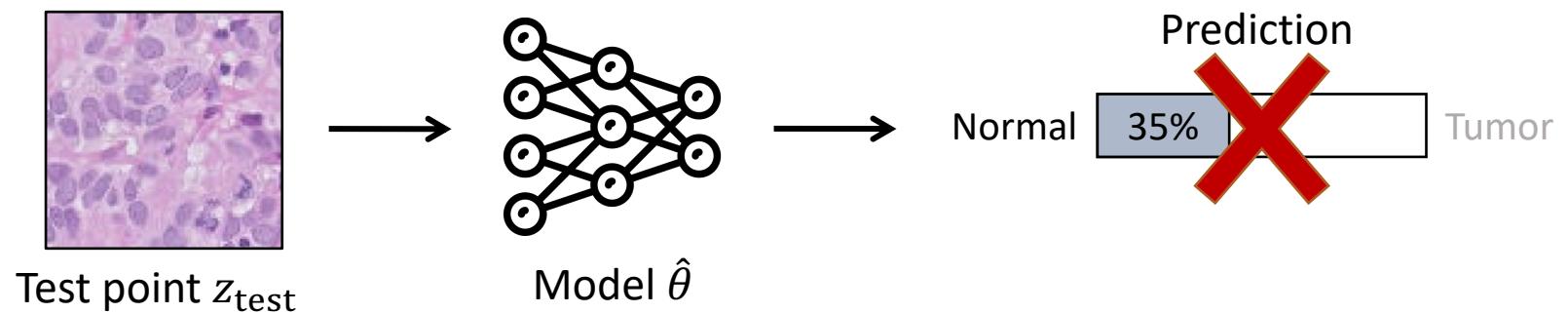


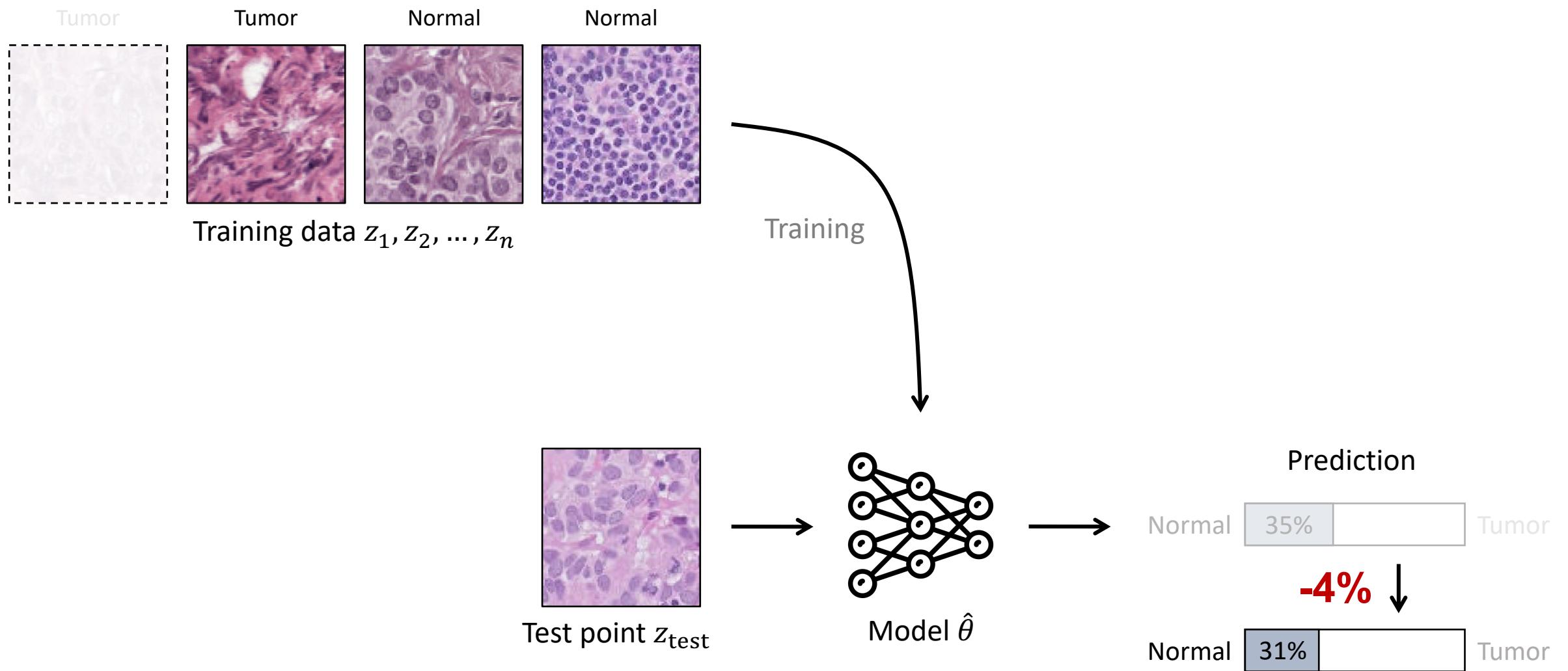


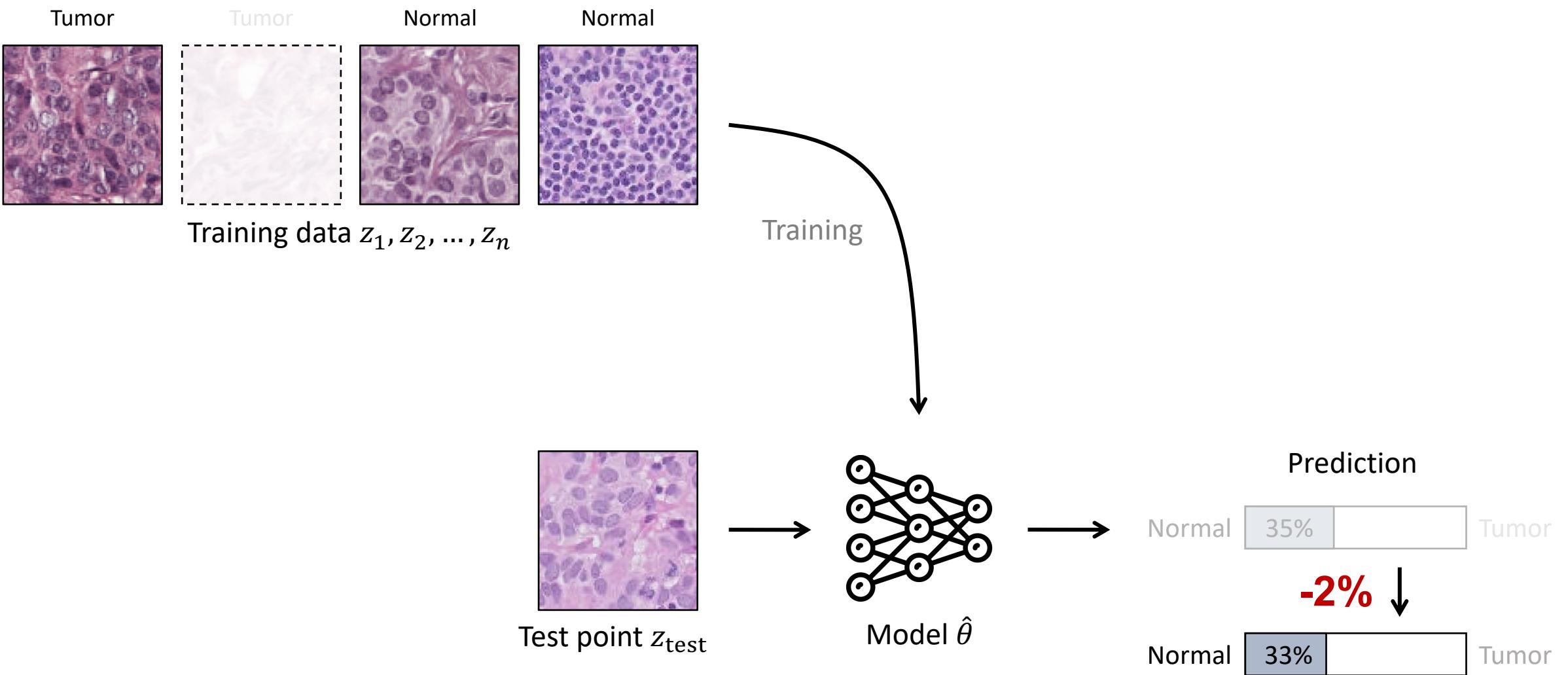


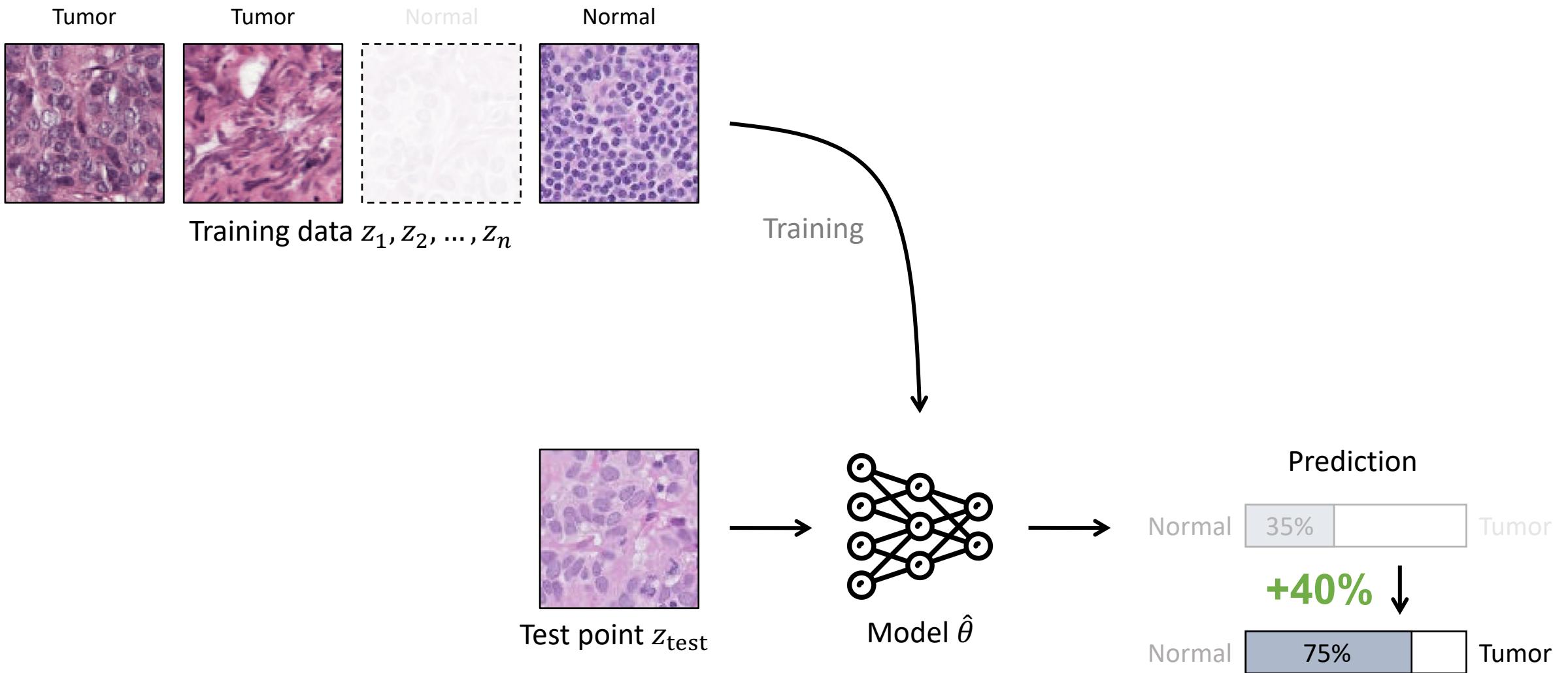


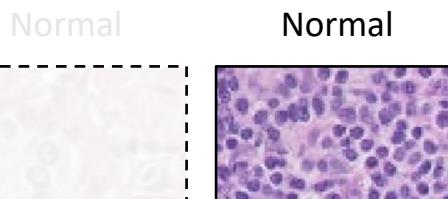
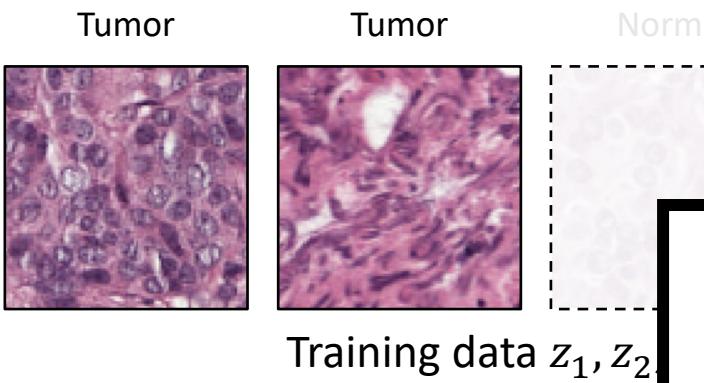
[Quenouille, 1956; Tukey, 1958]











Training data z_1, z_2

Problem

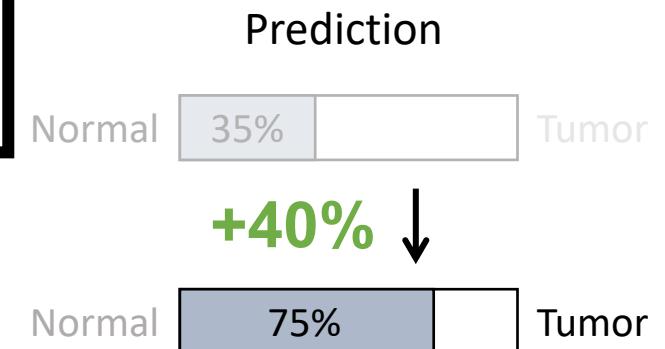
Repeatedly removing training
points and retraining is too slow

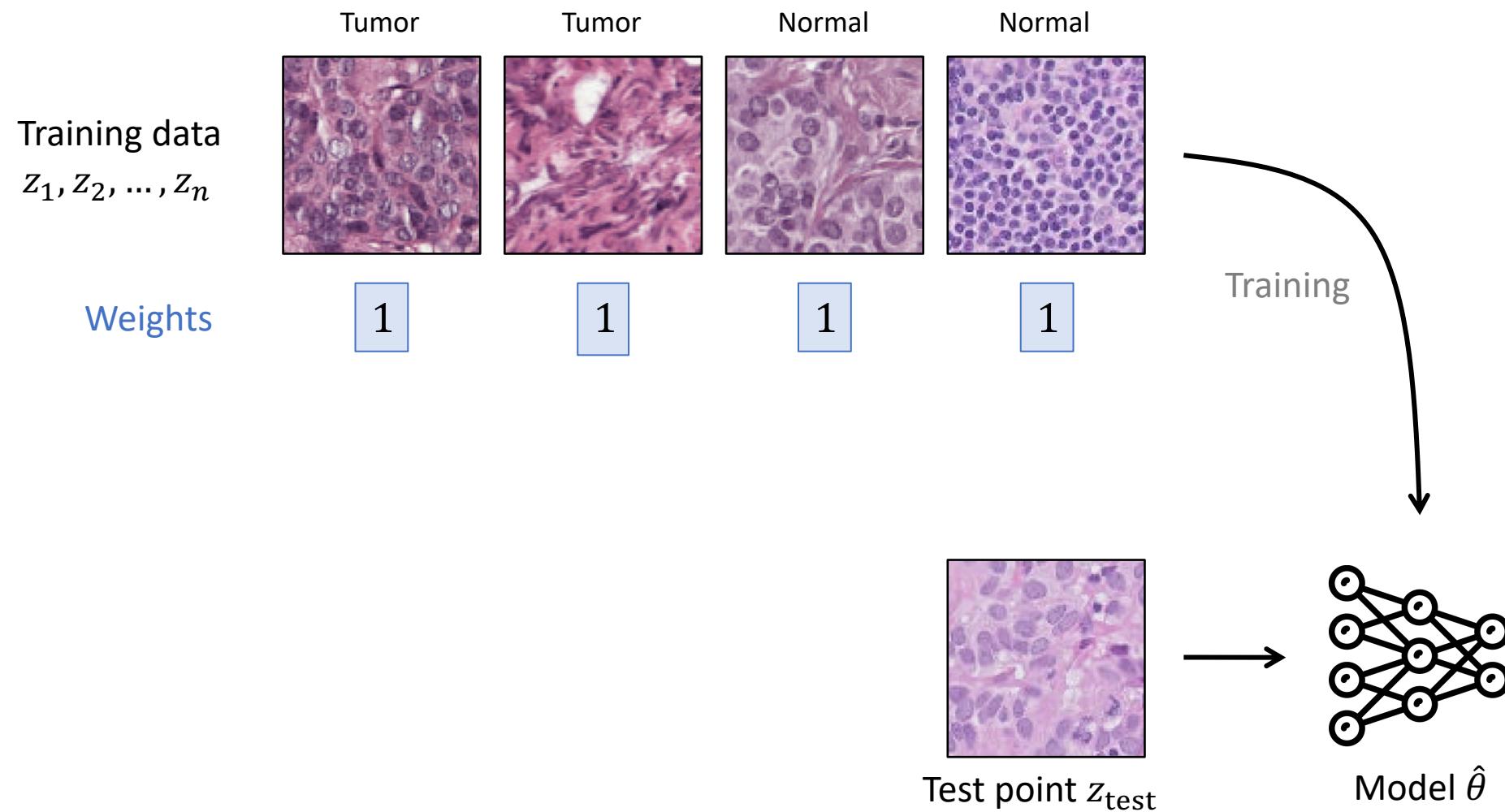
Solution

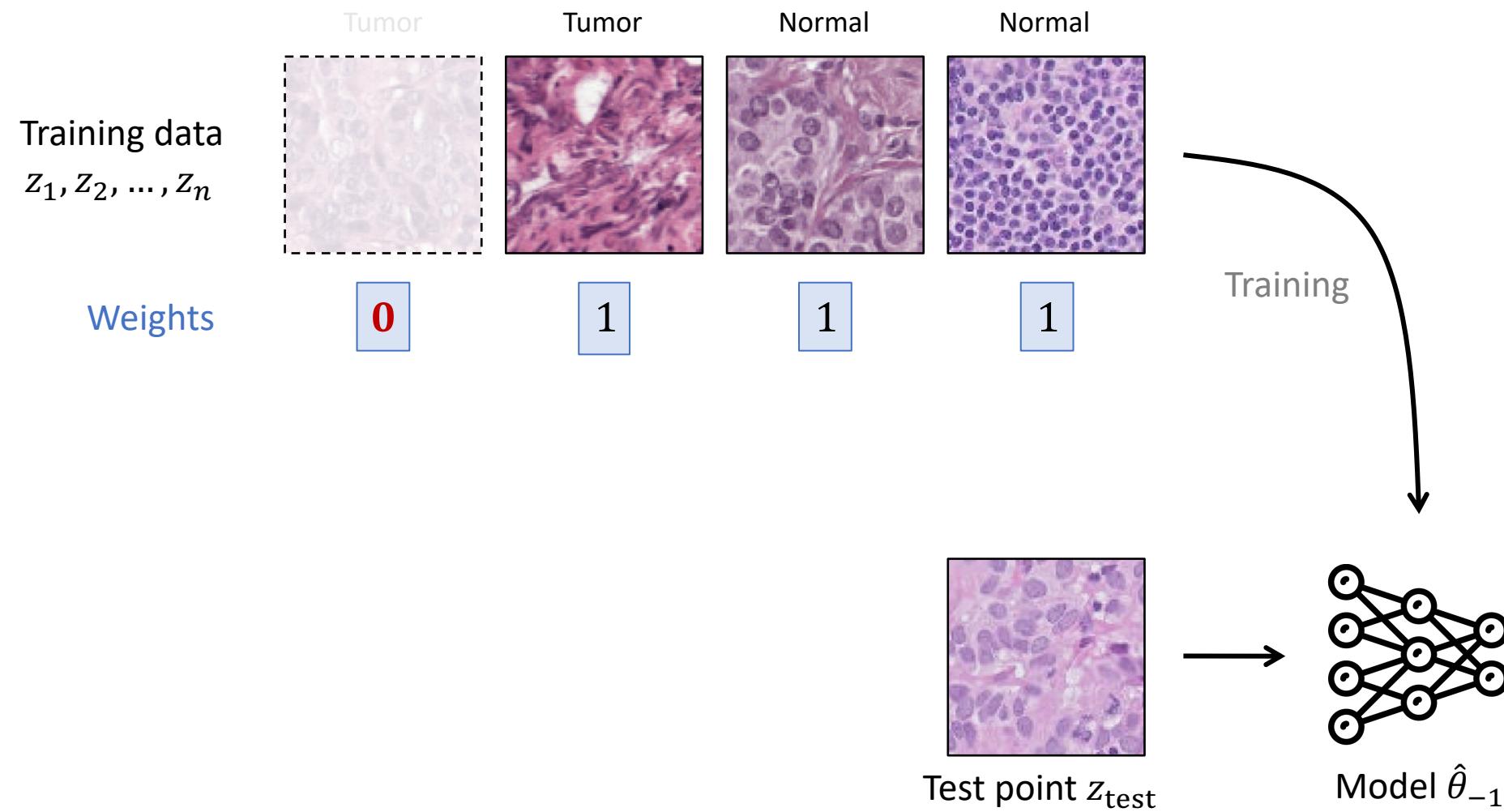
First-order Taylor approximation
via influence functions

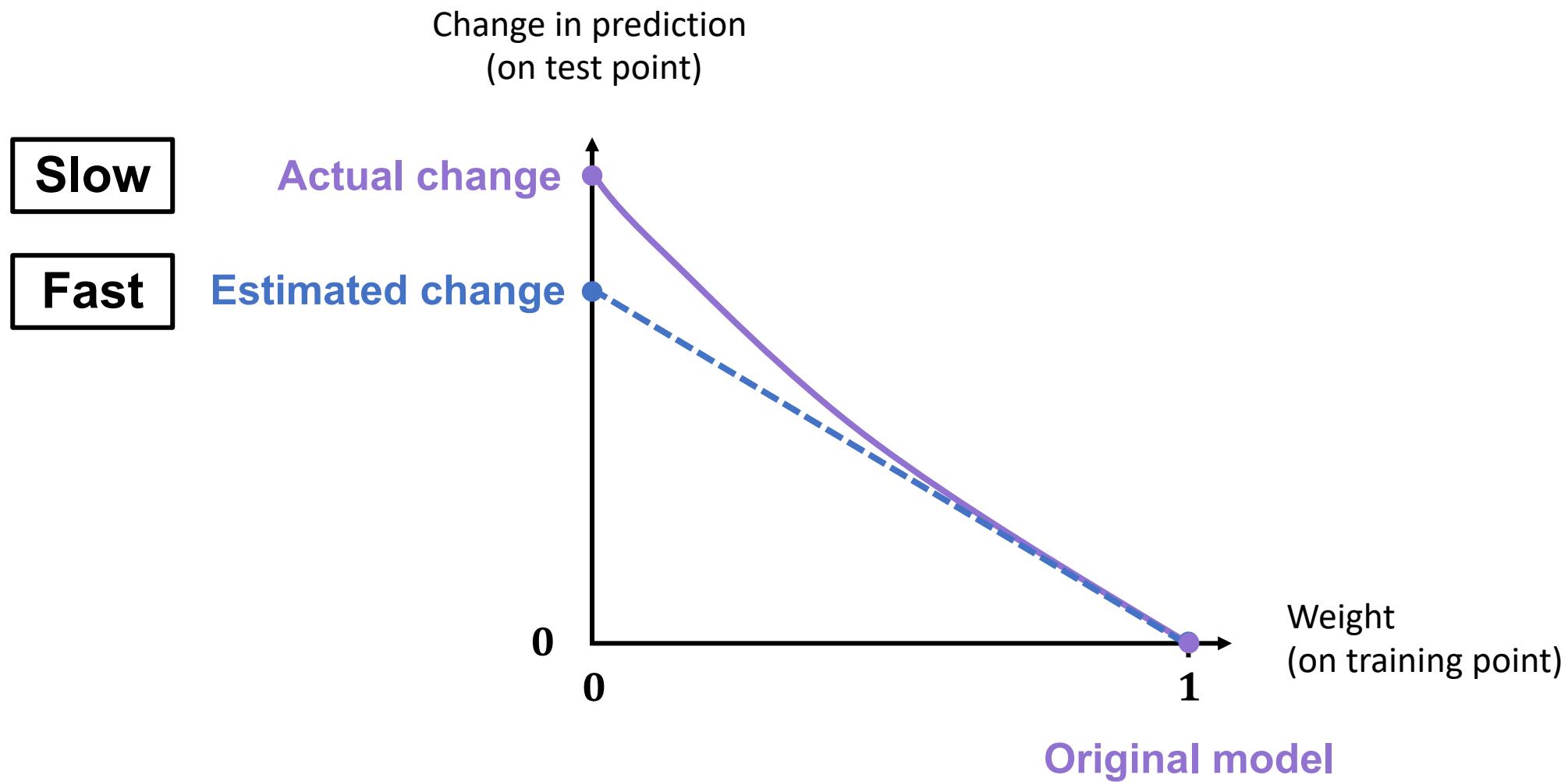


Model $\hat{\theta}$









The influence function approximation

$$\ell(z_{\text{test}}, \hat{\theta})$$

Loss on z_{test}
(original)

The influence function approximation

Change in loss on z_{test}
after removing z_{train}

$$\ell(z_{\text{test}}, \hat{\theta}_{-z_{\text{train}}}) - \ell(z_{\text{test}}, \hat{\theta})$$

Loss on z_{test}
(after removing z_{train}) Loss on z_{test}
(original)

The influence function approximation

Change in loss on z_{test}
after removing z_{train}

$$\ell(z_{\text{test}}, \hat{\theta}_{-z_{\text{train}}}) - \ell(z_{\text{test}}, \hat{\theta})$$

The influence function approximation

Change in loss on z_{test}
after removing z_{train}

$$\ell(z_{\text{test}}, \hat{\theta}_{-z_{\text{train}}}) - \ell(z_{\text{test}}, \hat{\theta}) \approx \nabla_{\theta} \ell(z_{\text{test}}, \hat{\theta})^T H_{\hat{\theta}}^{-1} \nabla_{\theta} \ell(z_{\text{train}}, \hat{\theta})$$

Gradient of loss on z_{test}

Gradient of loss on z_{train}

Inverse of the Hessian

$$H_{\hat{\theta}} = \frac{1}{n} \sum_{i=1}^n \nabla_{\theta}^2 \ell(z_i, \hat{\theta})$$

The influence function approximation

Doesn't require
retraining!

Change in loss on z_{test}
after removing z_{train}

$$\ell(z_{\text{test}}, \hat{\theta}_{-z_{\text{train}}}) - \ell(z_{\text{test}}, \hat{\theta}) \approx \nabla_{\theta} \ell(z_{\text{test}}, \hat{\theta})^T H_{\hat{\theta}}^{-1} \nabla_{\theta} \ell(z_{\text{train}}, \hat{\theta})$$

Gradient of
loss on z_{test}

Gradient of
loss on z_{train}

Inverse of the Hessian

$$H_{\hat{\theta}} = \frac{1}{n} \sum_{i=1}^n \nabla_{\theta}^2 \ell(z_i, \hat{\theta})$$

The influence function approximation

$$\ell(z_{\text{test}}, \hat{\theta}_{-z_{\text{train}}}) - \ell(z_{\text{test}}, \hat{\theta}) \approx \nabla_{\theta} \ell(z_{\text{test}}, \hat{\theta})^T H_{\hat{\theta}}^{-1} \nabla_{\theta} \ell(z_{\text{train}}, \hat{\theta})$$

Change in loss on z_{test} after removing z_{train}

Effect of other training points

Model's representation of z_{test}

Model's representation of z_{train}

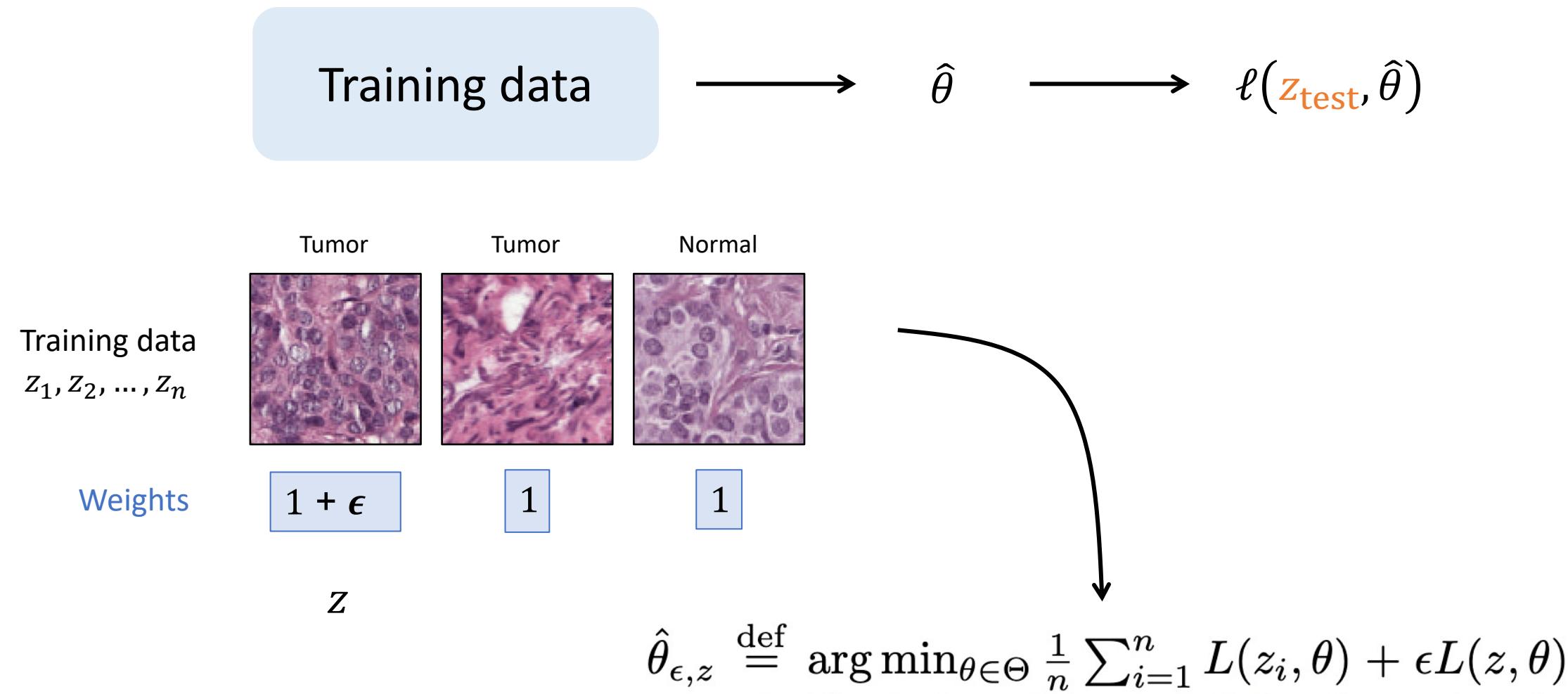
Gradient of loss on z_{test}

Gradient of loss on z_{train}

Inverse of the Hessian

$$H_{\hat{\theta}} = \frac{1}{n} \sum_{i=1}^n \nabla_{\theta}^2 \ell(z_i, \hat{\theta})$$

A little bit more technical details ...



A little bit more technical details ...

Training data

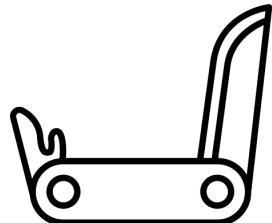
$$\hat{\theta} \longrightarrow \ell(z_{\text{test}}, \hat{\theta})$$

$$\frac{d\hat{\theta}_{\epsilon,z}}{d\epsilon} \Big|_{\epsilon=0} = -H_{\hat{\theta}}^{-1} \nabla_{\theta} L(z, \hat{\theta}),$$

$$\begin{aligned} \frac{d\hat{\theta}_{\epsilon,z}}{d\epsilon} \Big|_{\epsilon=0} & \quad \frac{dL(z_{\text{test}}, \hat{\theta}_{\epsilon,z})}{d\epsilon} \Big|_{\epsilon=0} \\ &= \nabla_{\theta} L(z_{\text{test}}, \hat{\theta})^{\top} \frac{d\hat{\theta}_{\epsilon,z}}{d\epsilon} \Big|_{\epsilon=0} \end{aligned}$$

$$\hat{\theta}_{\epsilon,z} \stackrel{\text{def}}{=} \arg \min_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^n L(z_i, \theta) + \epsilon L(z, \theta)$$

From classical to modern settings



Jaeckel, 1972. The infinitesimal jackknife.

Hampel, 1974. The influence curve and its role in robust estimation.

Cook, 1977. Detection of influential observations in linear regression.

...

From classical to modern settings

Small datasets
Low-dimensional \longrightarrow Large datasets
High-dimensional

Difficult to compute

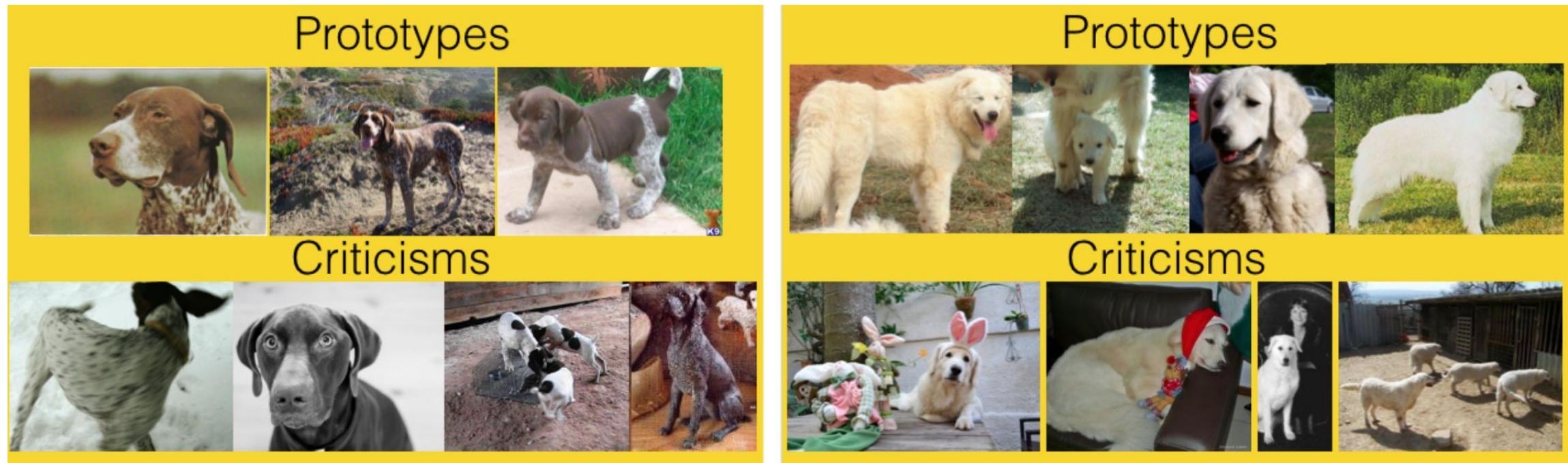
$$\nabla_{\theta} \ell(z_{\text{test}}, \hat{\theta})^T \boxed{H_{\hat{\theta}}^{-1}} \nabla_{\theta} \ell(z_{\text{train}}, \hat{\theta})$$

We use tools from
2nd-order optimization & stochastic estimation
[Pearlmutter, 1994; Martens, 2010, Agarwal et al., 2017]

Removing single points



Example based Explanations



Learned prototypes and criticisms from Imagenet dataset (two types of dog breeds)

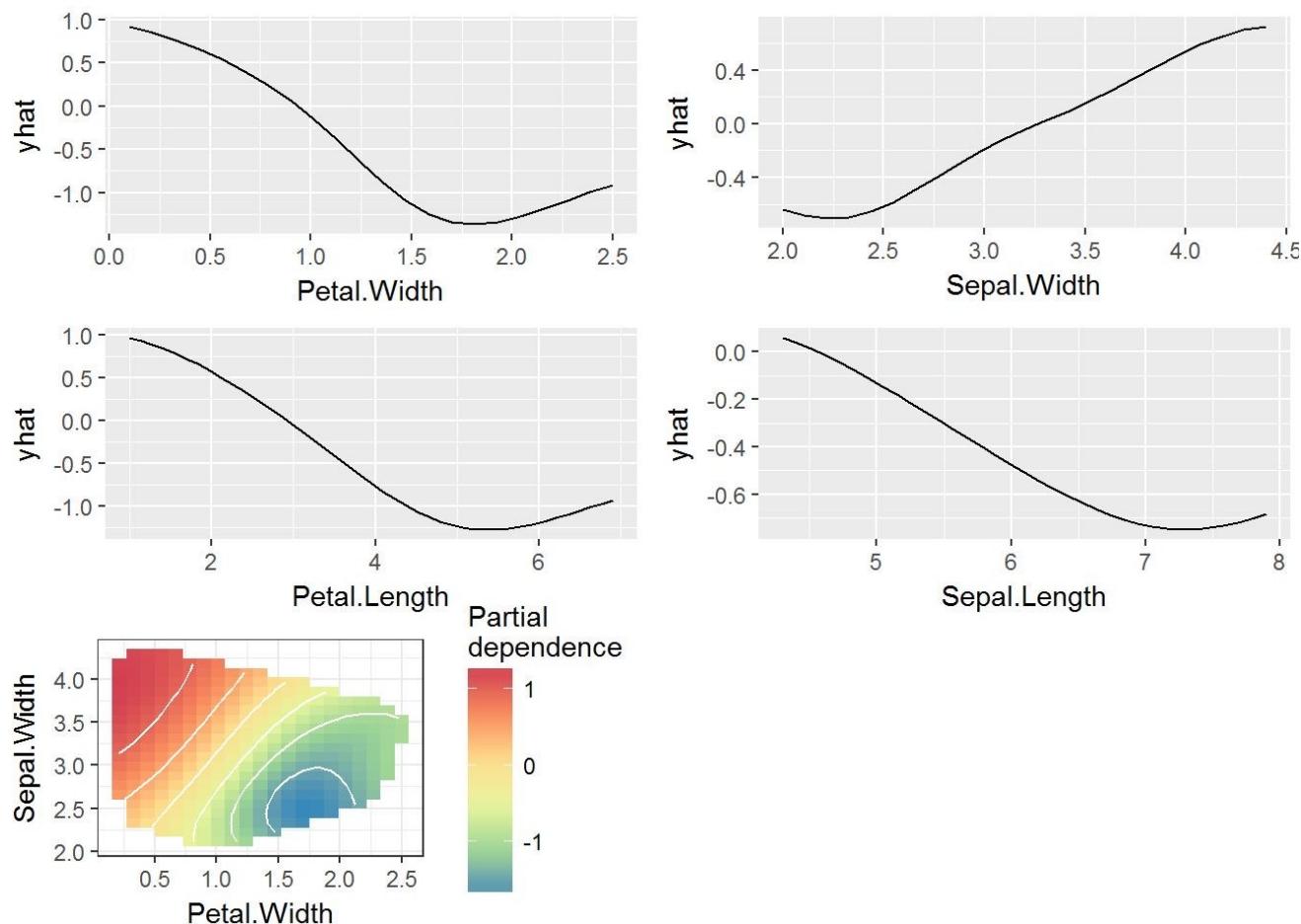
- Prototypes: Representative of all the training data.
- Criticisms: Data instance that is not well represented by the set of prototypes.



Global Explanations

Global Explanations Methods

- Partial Dependence Plot: Shows the marginal effect one or two features have on the predicted outcome of a machine learning model



Global Explanations Methods

- Permutations: The importance of a feature is the increase in the prediction error of the model after we permuted the feature's values, which breaks the relationship between the feature and the true outcome.

	RD Spend	Administration	Marketing Spend	Profit	state_California
1	165349.2	136897.8	471784.1	192261.83	0
2	162597.7	151377.59	443898.53	191792.06	1
3	153441.51	101145.55	407934.54	191050.39	1
...
48	0	135426.92	0	42559.73	1
49	542.05	51743.15	0	35673.41	0
50	0	116983.8	45173.06	14681.4	1

Random Shuffle of the first feature

Achieving Explainable AI

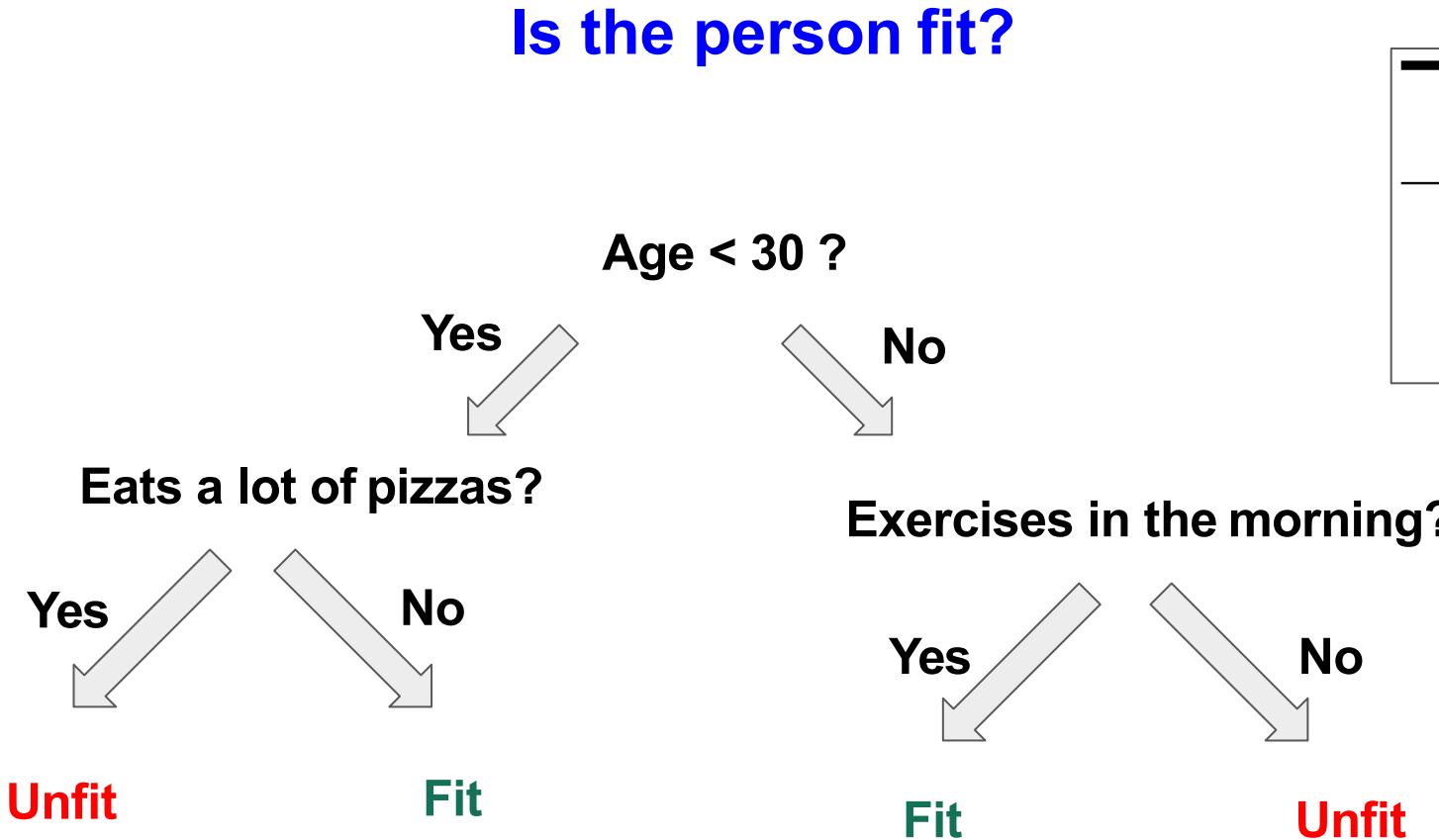
Approach 1: Post-hoc explain a given AI model

- Individual prediction explanations in terms of input features, influential examples, concepts, local decision rules
- Global prediction explanations in terms of entire model in terms of partial dependence plots, global feature importance, global decision rules

Approach 2: Build an interpretable model

- Logistic regression, Decision trees, Decision lists and sets, Generalized Additive Models (GAMs)

Decision Trees



Optimal Sparse Decision Trees

Xiyang Hu¹, Cynthia Rudin², Margo Seltzer^{3*}

¹Carnegie Mellon University, xiyanghu@cmu.edu

²Duke University, cynthia@cs.duke.edu

³The University of British Columbia, mseltzer@cs.ubc.ca

Decision Set

- If Allergies =Yes and Smoker =Yes and Irregular-Heartbeat =Yes, then Asthma
- If Allergies =Yes and Past-Respiratory-Illness =Yes and Avg-Body-Temperature ≥ 0.1 , then Asthma
- If Smoker =Yes and BMI ≥ 0.2 and Age ≥ 60 , then Diabetes
- If Family-Risk-Diabetes =Yes and BMI ≥ 0.4 =Frequency-Infections ≥ 0.2 , then Diabetes
- If Frequency-Doctor-Visits ≥ 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 ≥ 0.3 and Insurance-Coverage =None and Avg-Blood-Pressure ≥ 0.2 , then Depression
- If Past-Respiratory-Illness =Yes and Age ≥ 50 and Smoker =Yes, then Lung Cancer
- If Family-Risk-LungCancer =Yes and Allergies =Yes and Avg-Blood-Pressure ≥ 0.3 , then Lung Cancer
- If Disposition-Tiredness =Yes and Past-Anemia =Yes and BMI ≥ 0.3 and Rapid-Weight-Loss =Yes, then Leukemia
- If Family-Risk-Leukemia =Yes and Past-Blood-Clotting =Yes and Frequency-Doctor-Visits ≥ 0.3 , then Leukemia
- If Disposition-Tiredness =Yes and Irregular-Heartbeat =Yes and Short-Breath-Symptoms =Yes and Abdomen-Pains =Yes, then Myelofibrosis

Decision Set

A Bayesian Framework for Learning Rule Sets for Interpretable Classification

Tong Wang

TONG-WANG@UIOWA.EDU *University of Iowa*

Cynthia Rudin

CYNTHIA@CS.DUKE.EDU *Duke University*

Finale Doshi-Velez

FINALE@SEAS.HARVARD.EDU *Harvard University*

Yimin Liu

LIUYIMIN2000@GMAIL.COM *Edward Jones*

Erica Klampfl

EKLAMPFL@FORD.COM *Ford Motor Company*

Perry MacNeille

PMACNEIL@FORD.COM *Ford Motor Company*

Editor: Maya Gupta

Abstract

We present a machine learning algorithm for building classifiers that are comprised of a *small* number of *short* rules. These are restricted disjunctive normal form models. An example of a classifier of this form is as follows: *If X satisfies (condition A AND condition B) OR (condition C) OR ... , then Y = 1.* Models of this form have the advantage of being interpretable to human experts

since they produce a set of rules that concisely describe a specific class. We present two probabilistic models with prior parameters that the user can set to encourage the model to have a desired size and shape, to conform with a domain-specific definition of interpretability. We provide a scalable MAP inference approach and develop theoretical bounds to reduce computation by iteratively pruning the search space. We apply our method (Bayesian Rule Sets – *BRS*) to characterize and predict user behavior with respect to in-vehicle context-aware personalized recommender systems. Our method has a major advantage over classical associative classification methods and decision trees in that it does not greedily grow the model.

Decision List

```
If Past-Respiratory-Illness =Yes and Smoker =Yes and Age ≥ 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 ≥ 0.2 and Age≥ 60, then Diabetes  
Else if Frequent-Headaches =Yes and Dizziness =Yes, then Depression  
Else if Frequency-Doctor-Visits ≥ 0.3, then Diabetes  
Else if Disposition-Tiredness =Yes, then Depression  
Else if Chest-Pain =Yes and Nausea and Yes, then Diabetes  
Else Diabetes
```

Falling Rule List

A falling rule list is an ordered list of if-then rules (falling rule lists are a type of decision list), such that the estimated probability of success decreases monotonically down the list. Thus, a falling rule list directly contains the decision-making process, whereby the most at-risk observations are classified first, then the second set, and so on.

	Conditions		Probability	Support
IF	IrregularShape AND Age \geq 60	THEN malignancy risk is	85.22%	230
ELSE IF	SpiculatedMargin AND Age \geq 45	THEN malignancy risk is	78.13%	64
ELSE IF	IllDefinedMargin AND Age \geq 60	THEN malignancy risk is	69.23%	39
ELSE IF	IrregularShape	THEN malignancy risk is	63.40%	153
ELSE IF	LobularShape AND Density \geq 2	THEN malignancy risk is	39.68%	63
ELSE IF	RoundShape AND Age \geq 60	THEN malignancy risk is	26.09%	46
ELSE		THEN malignancy risk is	10.38%	366

Falling rule list for mammographic mass dataset.

Box Drawings for Rare Classes

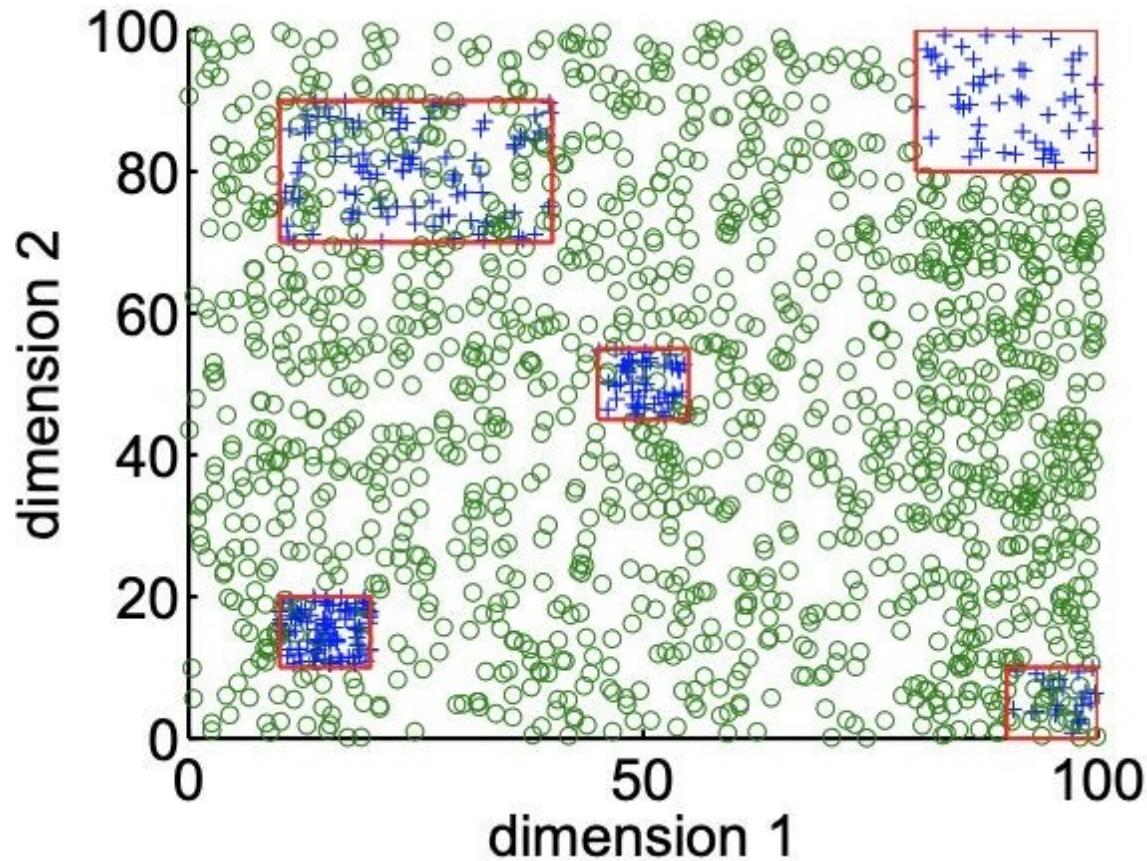


Figure credit: Box Drawings for Learning with Imbalanced. Data Siong Thye Goh and Cynthia Rudin

Supersparse Linear Integer Models for Optimized Medical Scoring Systems

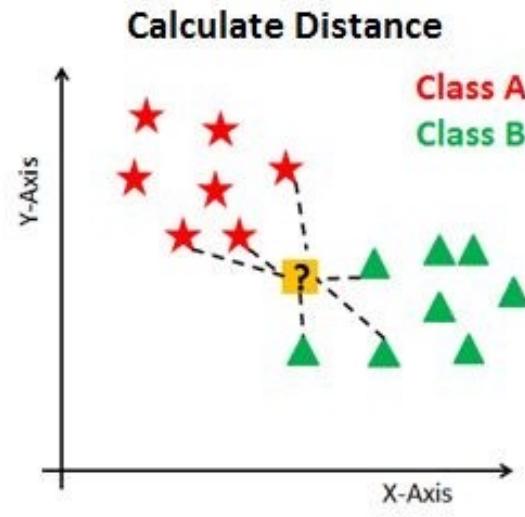
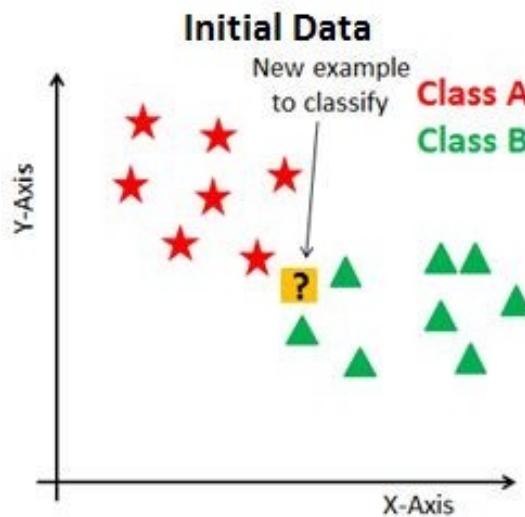
PREDICT PATIENT HAS OBSTRUCTIVE SLEEP APNEA IF SCORE > 1

1. <i>age</i> ≥ 60	4 points
2. <i>hypertension</i>	4 points	+
3. <i>body mass index</i> ≥ 30	2 points	+
4. <i>body mass index</i> ≥ 40	2 points	+
5. <i>female</i>	-6 points	+
ADD POINTS FROM ROWS 1 – 5	SCORE	=

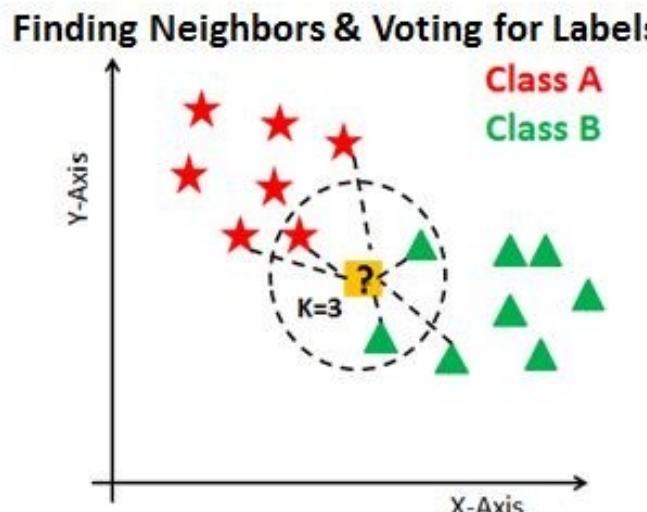
SLIM scoring system for sleep apnea screening. This model achieves a 10-CV mean test TPR/FPR of 61.4/20.9%, obeys all operational constraints, and was trained without parameter tuning. It also generalizes well due to the simplicity of the hypothesis space: here the training TPR/FPR of the final model is 62.0/19.6%.

Figure credit: Supersparse Linear Integer Models for Optimized Medical Scoring Systems. Berk Ustun and Cynthia Rudin

K- Nearest Neighbors



Explanation in terms of nearest training data points responsible for the decision



GLMs and GAMs

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

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



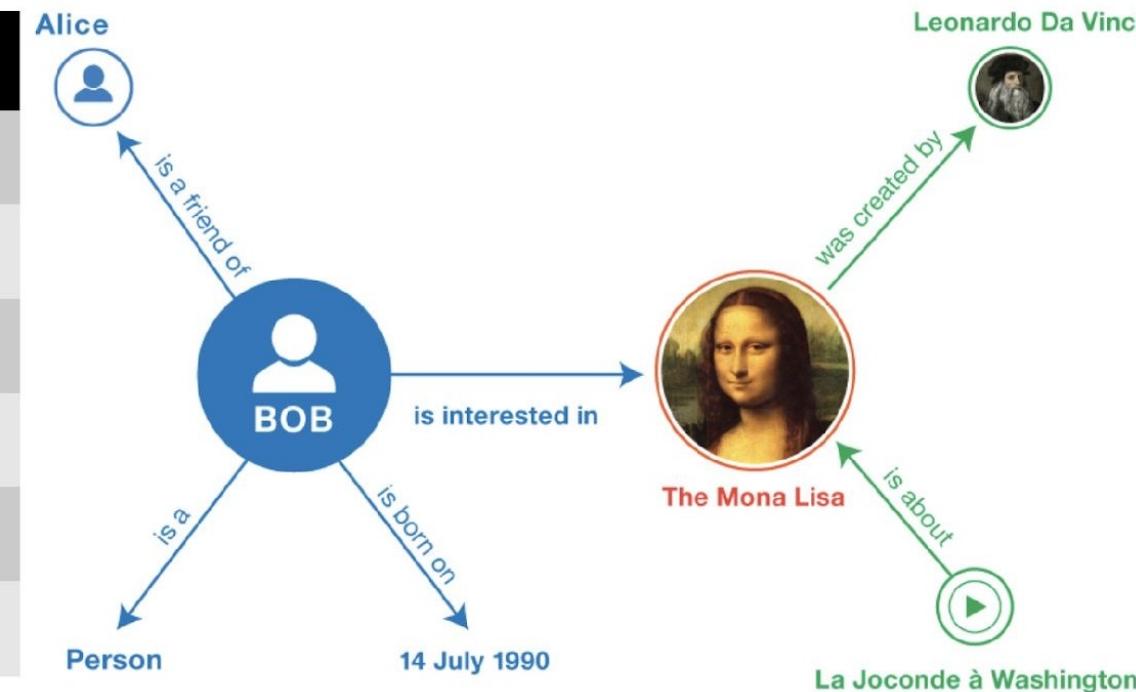
Explainable Machine Learning

(from a Knowledge
Graph Perspective)

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	<i>is interested in</i>	The Mona Lisa
Bob	<i>is a friend of</i>	Alice
The Mona Lisa	<i>was created by</i>	Leonardo Da Vinci
Bob	<i>is a</i>	Person
La Joconde à W.	<i>is about</i>	The Mona Lisa
Bob	<i>is born on</i>	14 July 1990



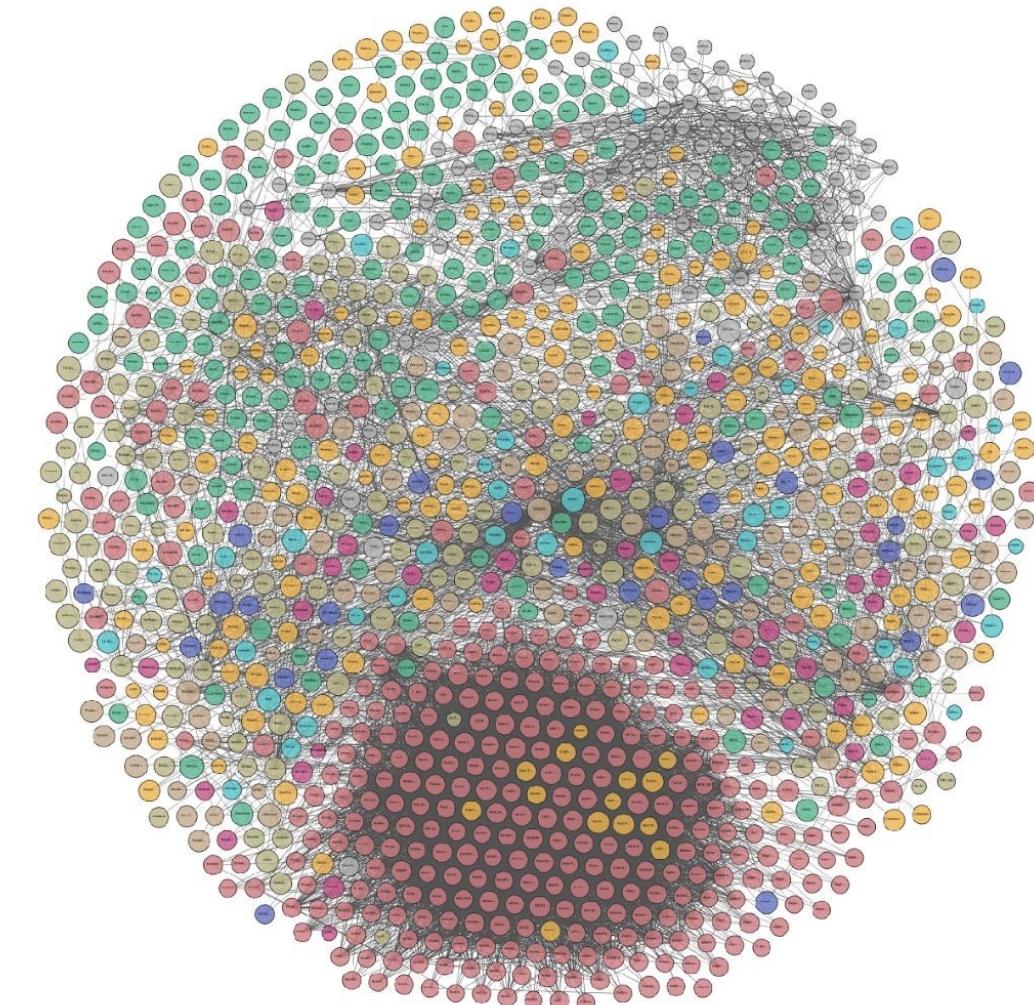
Knowledge Graph (2)

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

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

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

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



Knowledge Graph Construction

Knowledge Graph construction methods can be classified in:

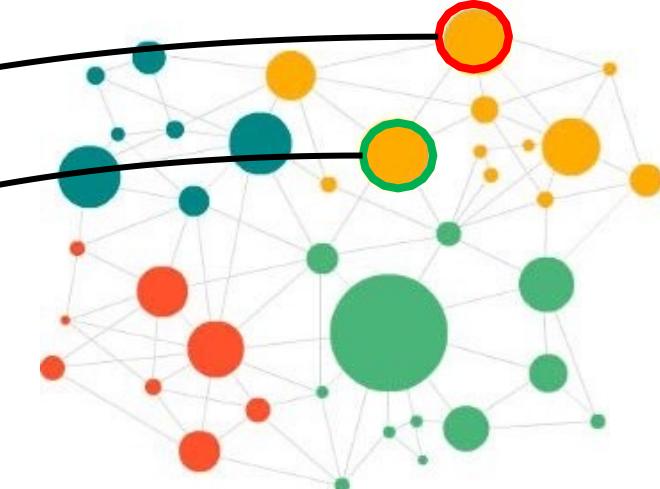
- **Manual** – curated (e.g. via experts), collaborative (e.g. via volunteers)
- **Automated** – semi-structured (e.g. from infoboxes), unstructured (e.g. from text)

Coverage is an issue:

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

Relational Learning can help us overcoming these issues.

Knowledge Graph in Machine Learning (1)

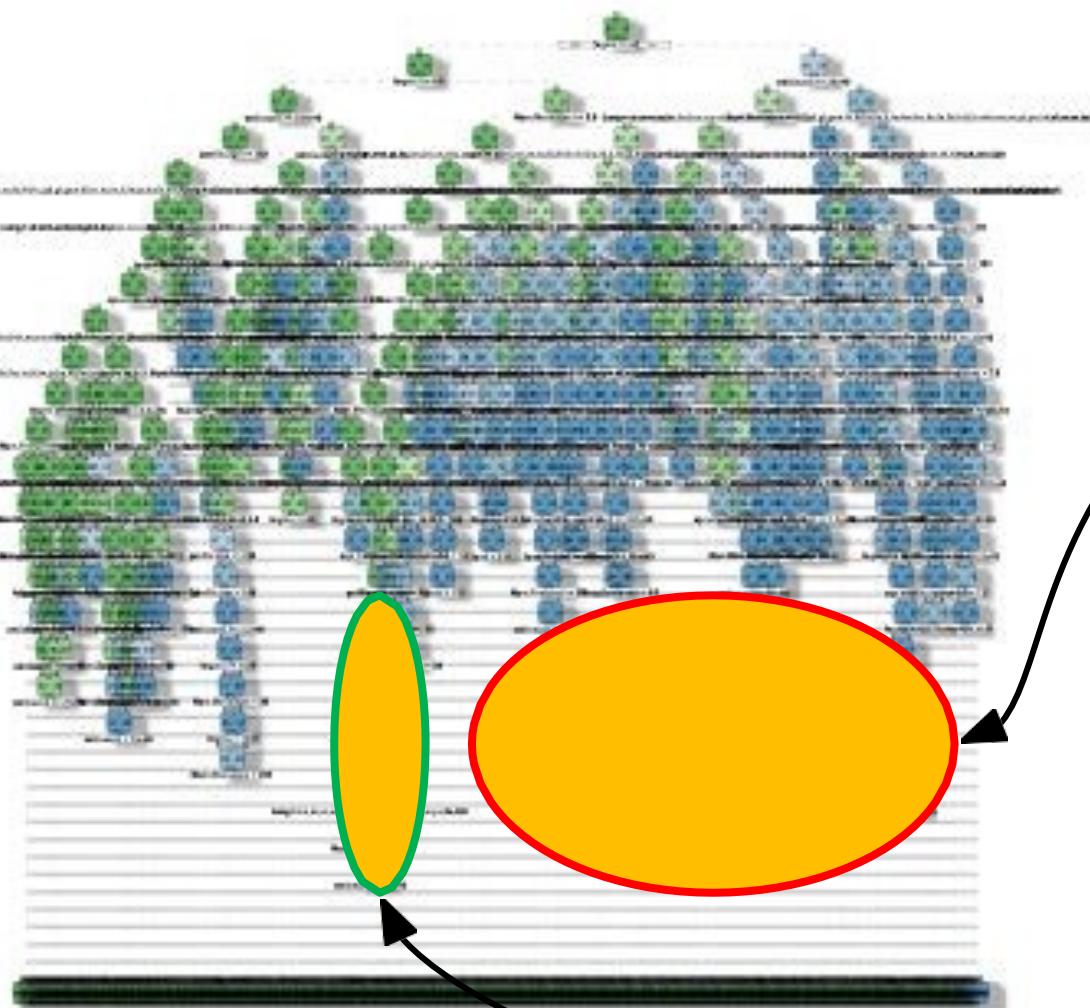


Augmenting (input) features
with more semantics such as
knowledge graph embeddings /
entities

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

Freddy Lécué: On the role of knowledge graphs in
explainable AI. Semantic Web 11(1): 41-51 (2020)
158

Knowledge Graph in Machine Learning (2)

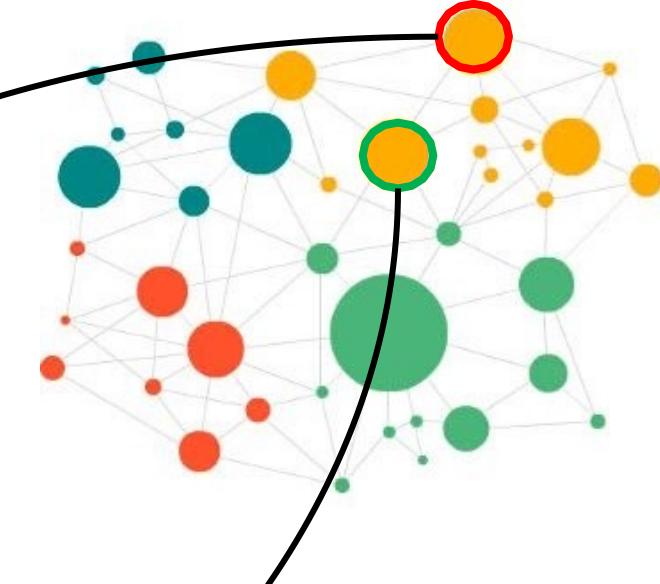


Rattle 2016-Aug-18 16:15:42 sklisarov

<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)
159



Knowledge Graph in Machine Learning (3)

● Input Layer

Training Data

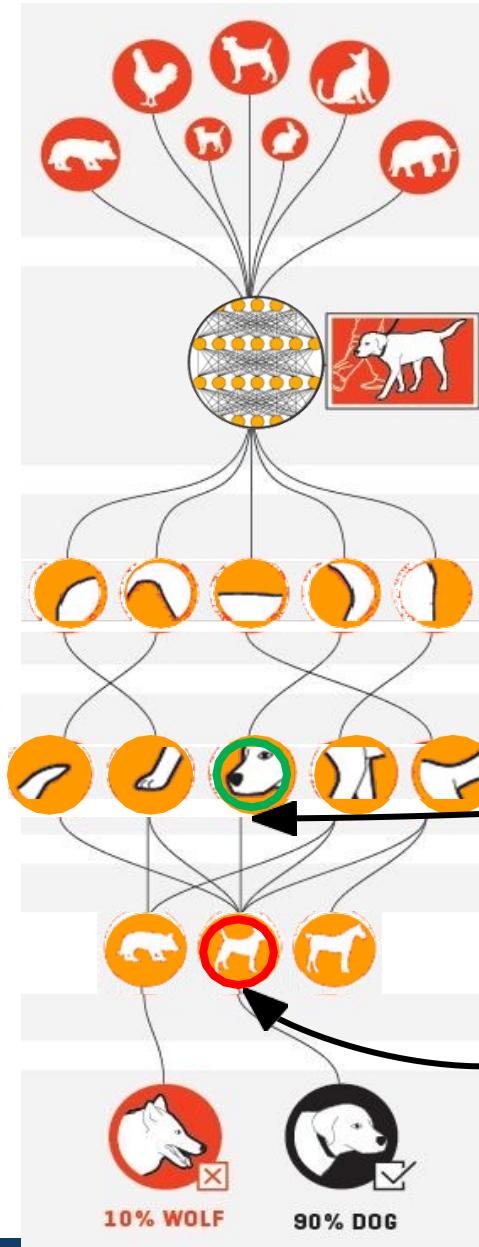
○ Hidden Layer

Neurons respond to simple shapes

Neurons respond to more complex structures

Neurons respond to highly complex, abstract concepts

● Output Layer



Input
(unlabeled
image)

1st Layer

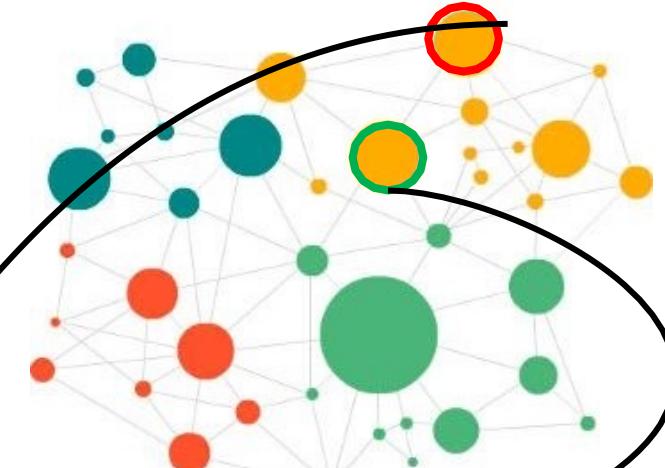
2nd Layer

nth Layer

Low-level
features to
high-level
features

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

Freddy Lécué: On the role of knowledge graphs in
explainable AI. Semantic Web 11(1): 41-51 (2020)
160



Knowledge Graph in Machine Learning (4)

● Input Layer

Training Data

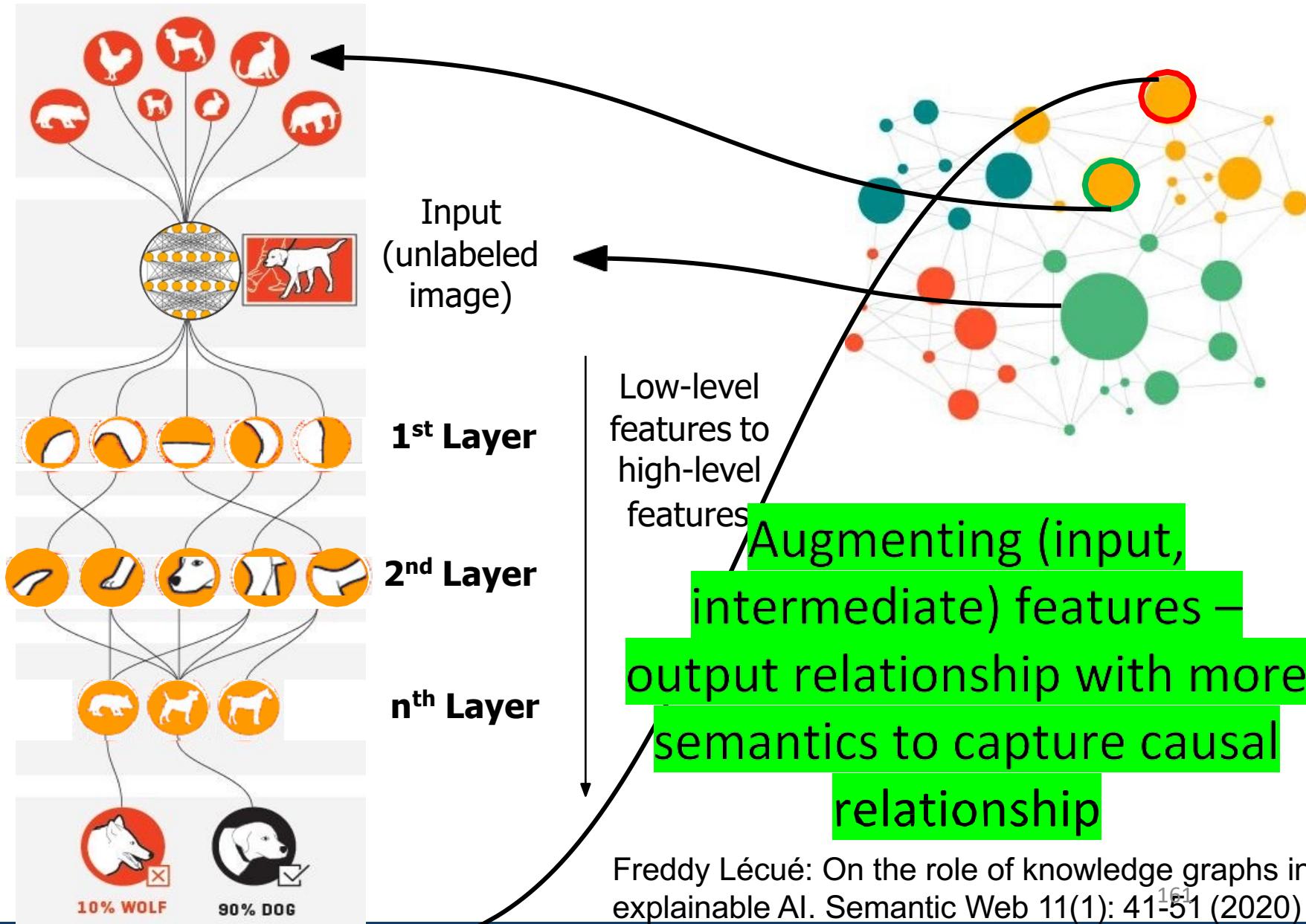
○ Hidden Layer

Neurons respond to simple shapes

Neurons respond to more complex structures

Neurons respond to highly complex, abstract concepts

● Output Layer



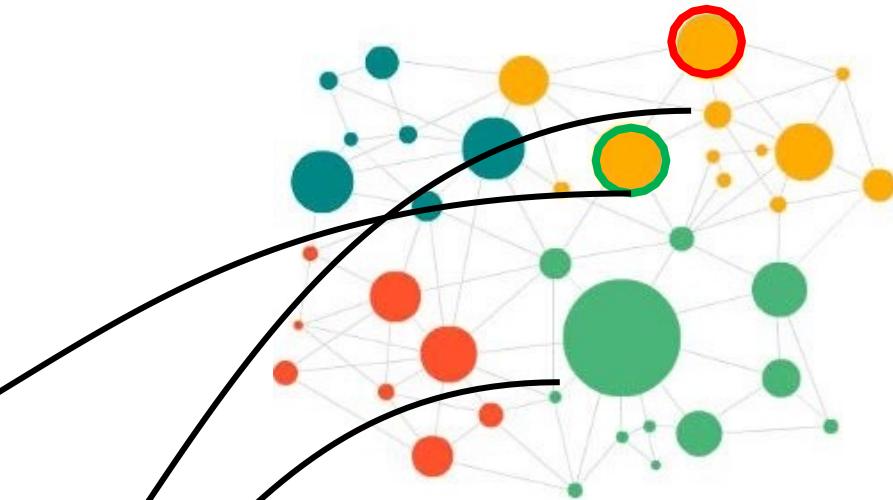
Knowledge Graph in Machine Learning (5)



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

Knowledge Graph in Machine Learning (6)

“How to explain transfer learning with appropriate knowledge representation?

Augmenting input features and domains with semantics to support interpretable transfer

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

Knowledge-Based Transfer Learning Explanation

Jiaoyan Chen
Department of Computer Science
University of Oxford, UK

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

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

Freddy Lecue
INRIA, France
Accenture Labs, Ireland

Ian Horrocks
Department of Computer Science
University of Oxford, UK

How Does
it
Work
in Practice?

State of the Art Machine Learning Applied to Critical Systems

Object (Obstacle) Detection Task

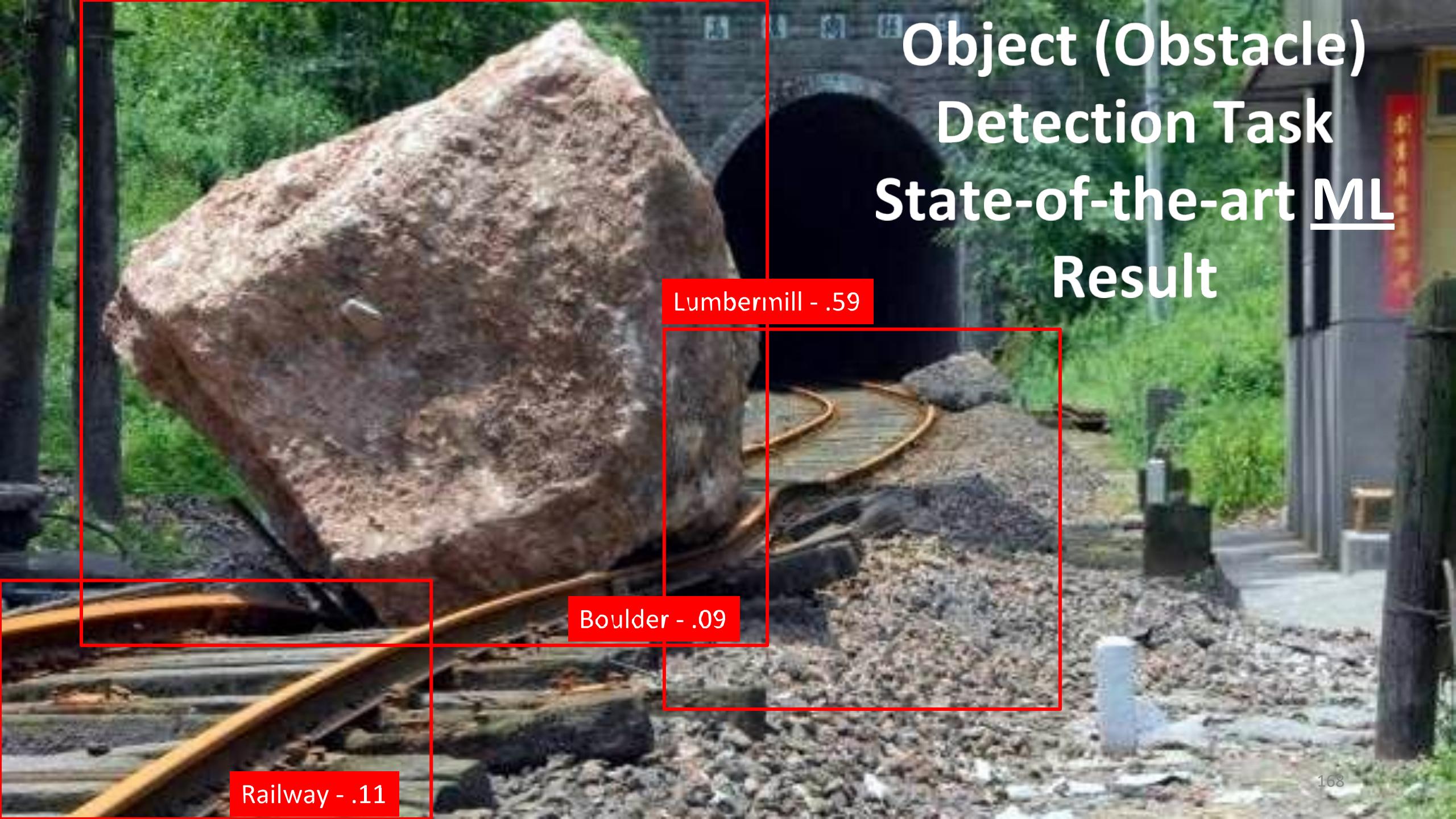


Object (Obstacle) Detection Task State-of-the-art ML Result

Lumbermill - .59



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



State of the Art

XAI

**Applied to Critical
Systems**

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



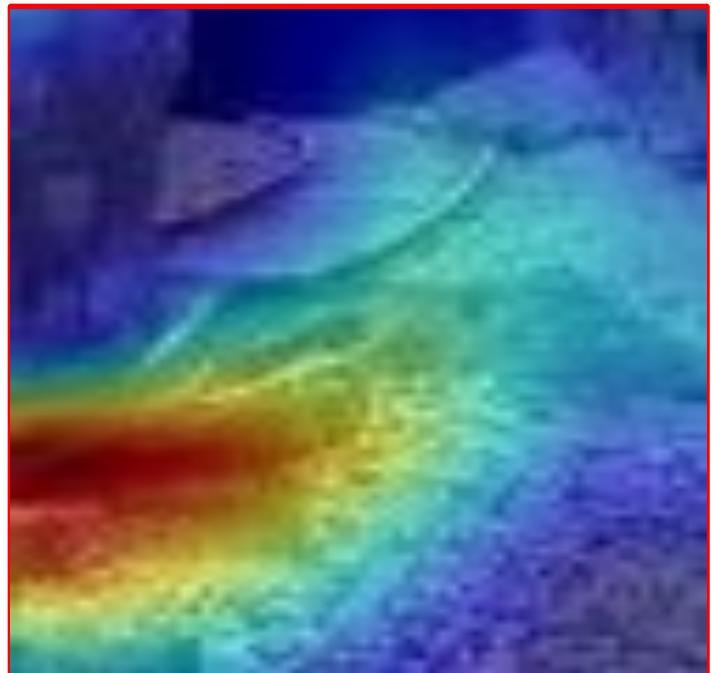
**Unfortunately, this is of
NO use for a human
behind the system**

Let's stay back

Why this Explanation? (meta explanation)

After Human Reasoning...

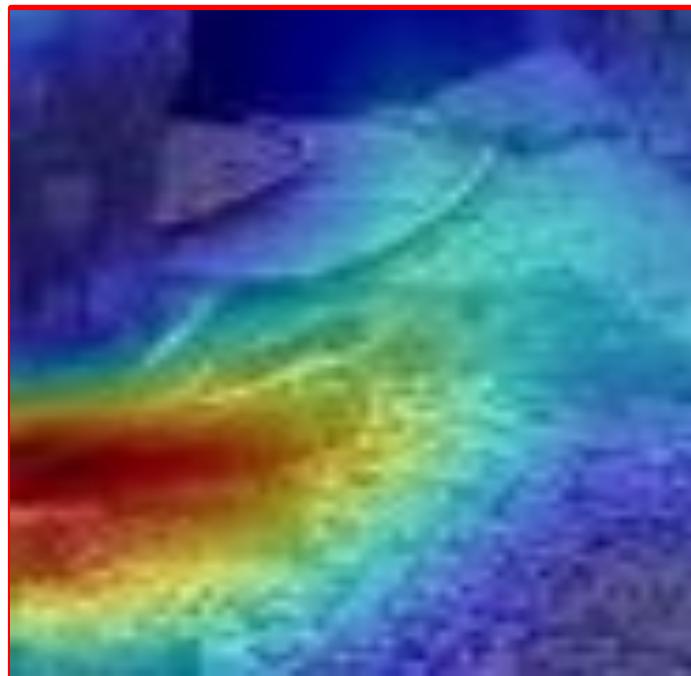
Lumbermill - .59



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?

Lumbermill - .59



Context matters

Railway - .11

Boulder - .09

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 bollersten. Boulder sized clasts are found in some sedimentary rocks, such as coarse conglomerate and boulder clay.

Property	Value
dbo:abstract	<ul style="list-style-type: none">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 bollersten. 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	<ul style="list-style-type: none">wiki-commons:Special:FilePath/Balanced_Rock.jpg?width=300
dbo:wikiPageID	<ul style="list-style-type: none">60784 (xsd:integer)
dbo:wikiPageRevisionID	<ul style="list-style-type: none">743049914 (xsd:integer)
dc:tsubject	<ul style="list-style-type: none">dbc:Rock_formationsdbc:Rocks

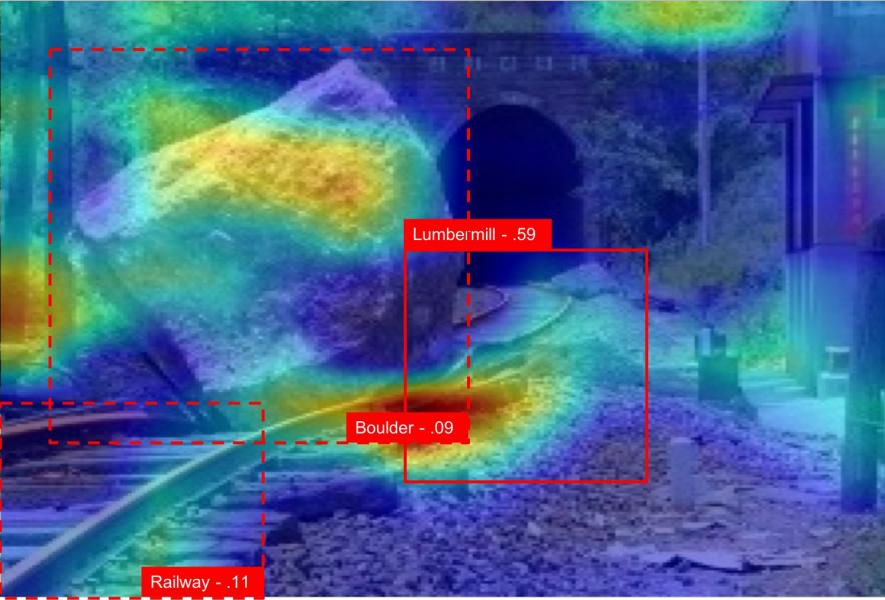
DBpedia [Browse using](#) [Formats](#) [Faceted Browser](#) [Sparql Endpoint](#)

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	<ul style="list-style-type: none">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,



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**

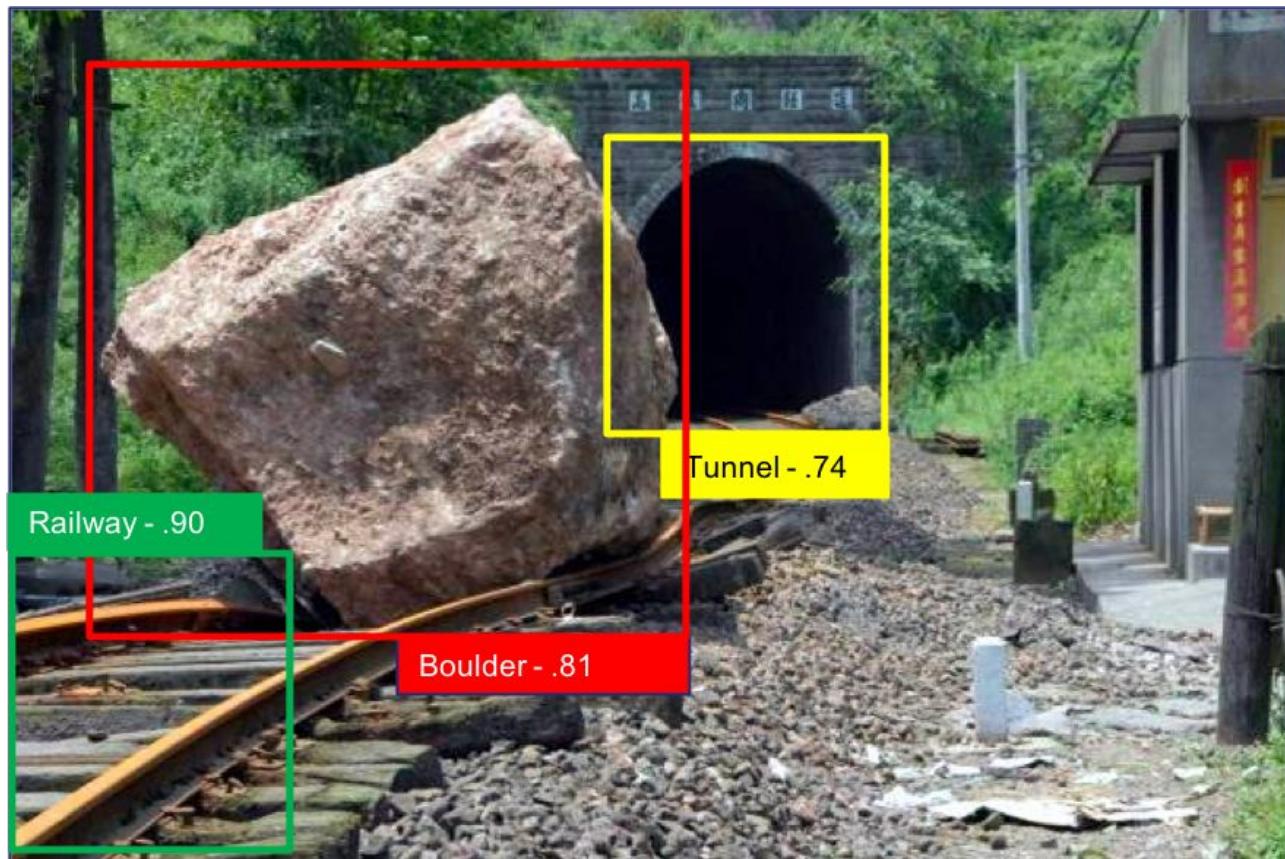


Transitioning



- **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

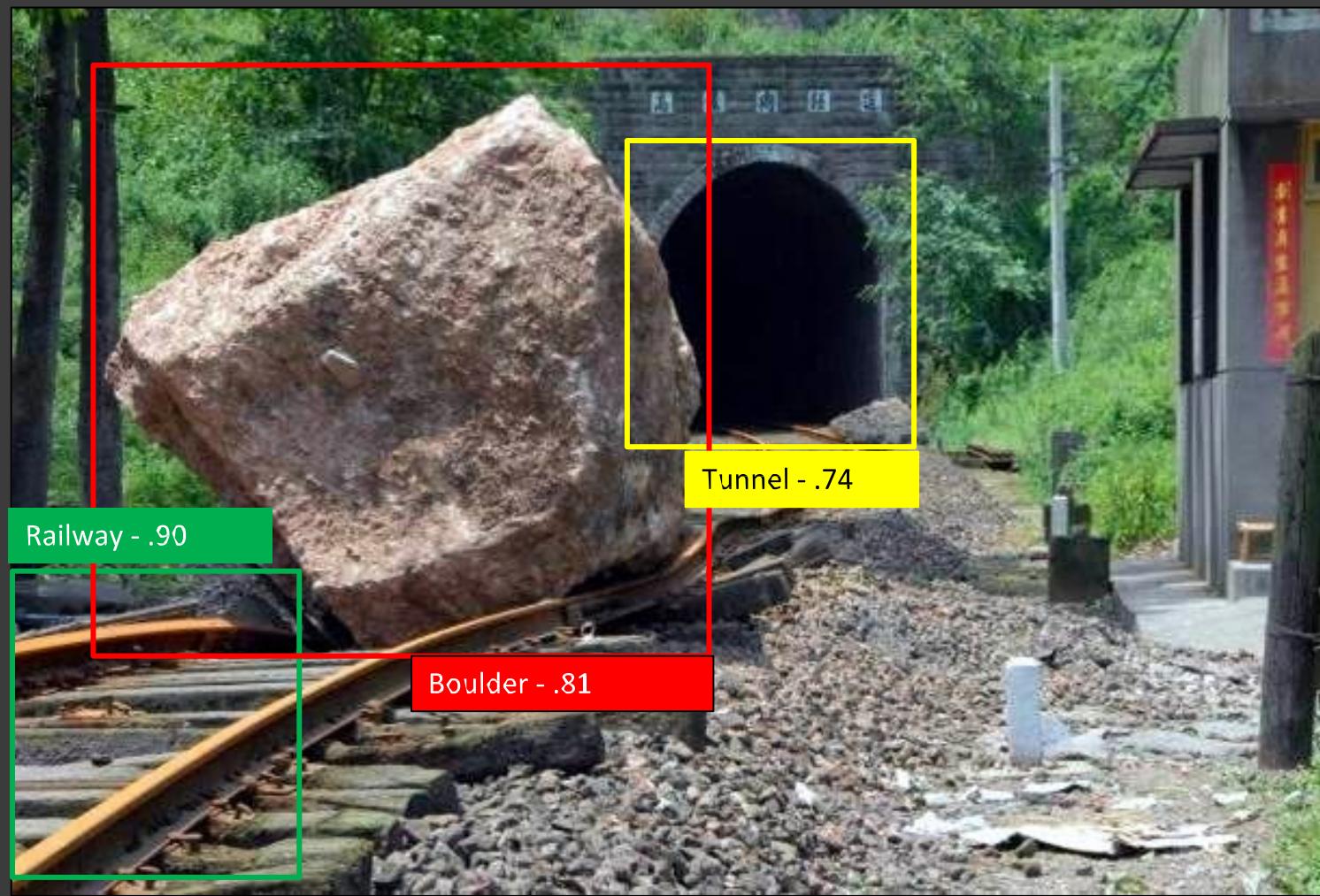
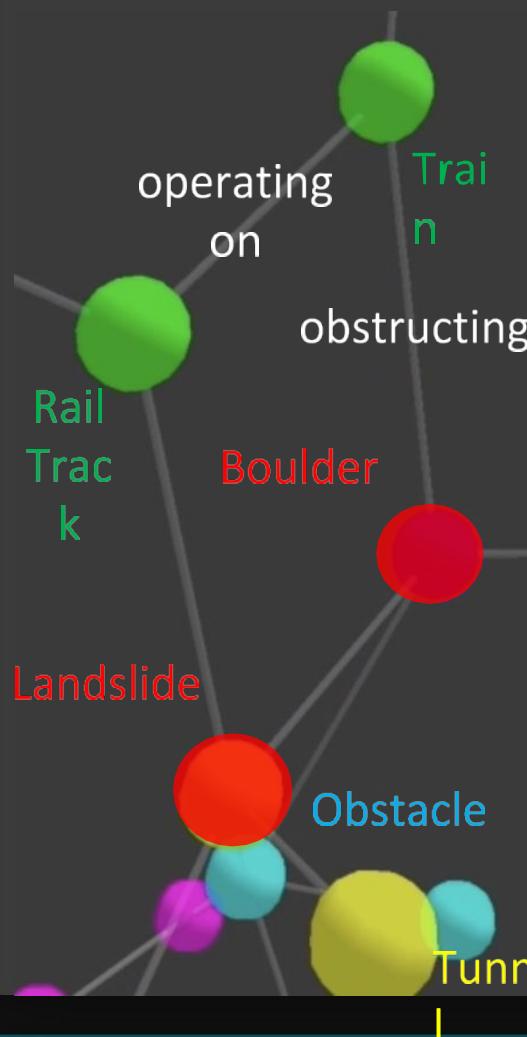


EXPLANATIONS

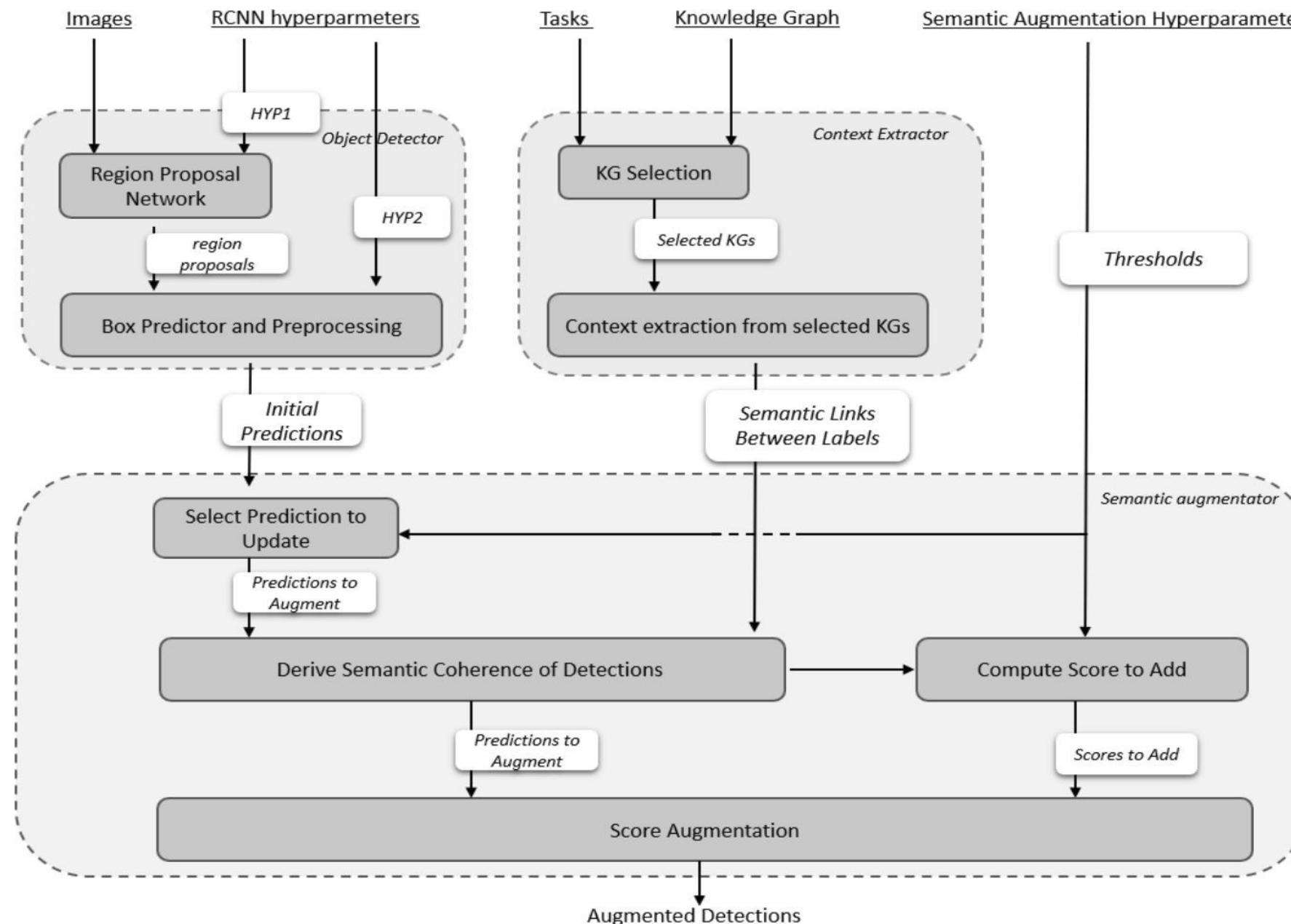
ResNet50 image classifier



Lime



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

Thank you!