AAAI 2020 Tutorial

Freddy Lecue, Krishna Gade, Sahin Cem Geyik, Krishnaram Kenthapadi, Varun Mithal, Ankur Taly, Riccardo Guidotti, Pasquale Minervini

https://xaitutorial2020.github.io

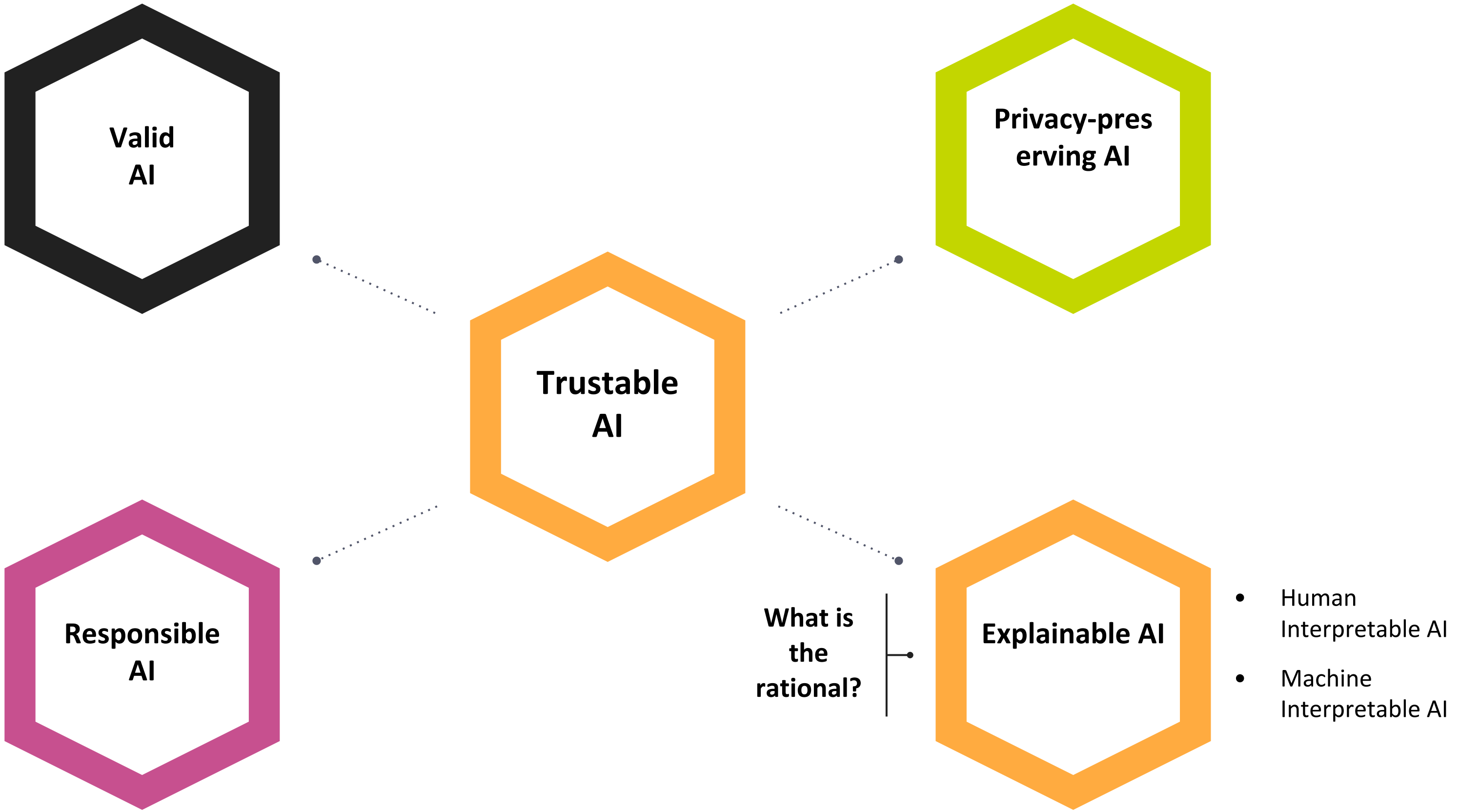
Outline

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)
- Part V: Case Studies from Industry
  - Applications, Lessons Learned, and Research Challenges

Scope

AI Adoption: Requirements



Introduction and Motivation

Explanation - From a Business Perspective

Business to Customer AI

Critical Systems (1)

Critical Systems (2)

… but not only Critical Systems (1)

COMPAS recidivism black bias


Opinion

OP-ED CONTRIBUTOR

When a Computer Program Keeps You in Jail

By Rebecca Wexler

June 13, 2017

… but not only Critical Systems (2)

Finance:

- Credit scoring, loan approval
- Insurance quotes

FICO COMMUNITY

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

The Big Read Artificial intelligence + Add to myFT

Insurance: Robots learn the business of covering risk

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

Save

Oliver Ralph MAY 16, 2017 24

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

… but not only Critical Systems (3)

Healthcare

- Applying ML methods in medical care is problematic.
- AI as 3rd-party actor in physician-patient relationship
- Responsibility, confidentiality?
- Learning must be done with available data.

  Cannot randomize cares given to patients!
- Must validate models before use.

Stanford MEDICINE | News Center

Email Tweet

Researchers say use of artificial intelligence in medicine raises ethical questions

In a perspective piece, Stanford researchers discuss the ethical implications of using machine-learning tools in making health care decisions for patients.

Patricia Hannon
,https://med.stanford.edu/news/all-news/2018/03/researchers-say-use-of-ai-in-medicine-raises-ethical-questions.html

Intelligible Models for HealthCare: Predicting Pneumonia Risk and Hospital 30-day Readmission

Rich Caruana
Microsoft Research
rcaruana@microsoft.com

Yin Lou
LinkedIn Corporation
ylou@linkedin.com

Johannes Gehrke
Microsoft
johannes@microsoft.com

Paul Koch
Microsoft Research
paulkoch@microsoft.com

Marc Sturm
NewYork-Presbyterian Hospital
mas9161@nyp.org

Noémie Elhadad
Columbia University
noemie.elhadad@columbia.edu

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

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

MIT News

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

Feb 12, 2018

BBC NEWS

Tay: Microsoft issues apology over racist chatbot fiasco

Sep 22, 2017

Missouri S&T News and Research

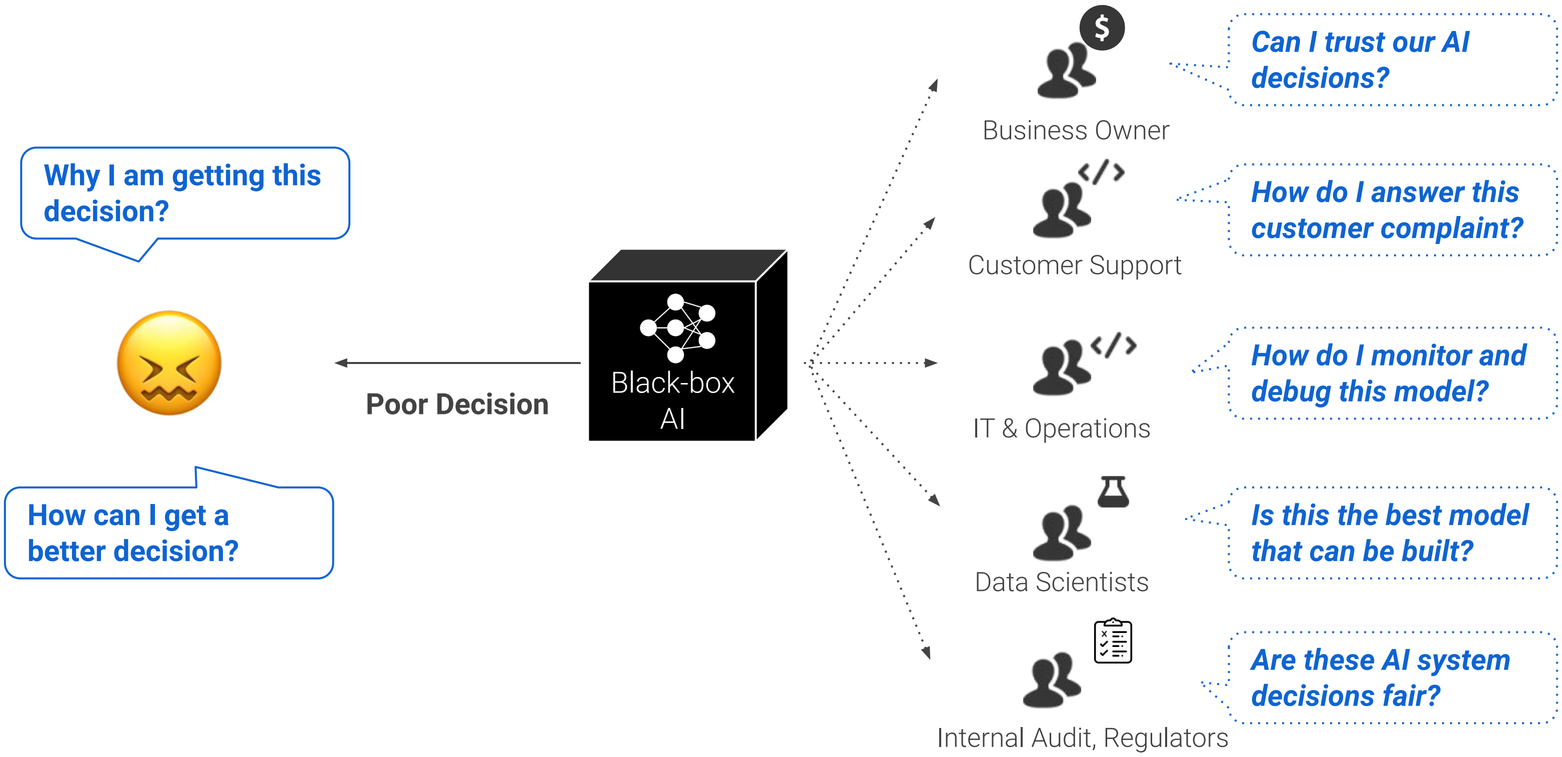
After Uber, Tesla incidents, can artificial intelligence be trusted?

Apr 10, 2018

Guilty! AI Is Found to Perpetuate Biases in Jailing

1 day ago

Black-box AI creates confusion and doubt



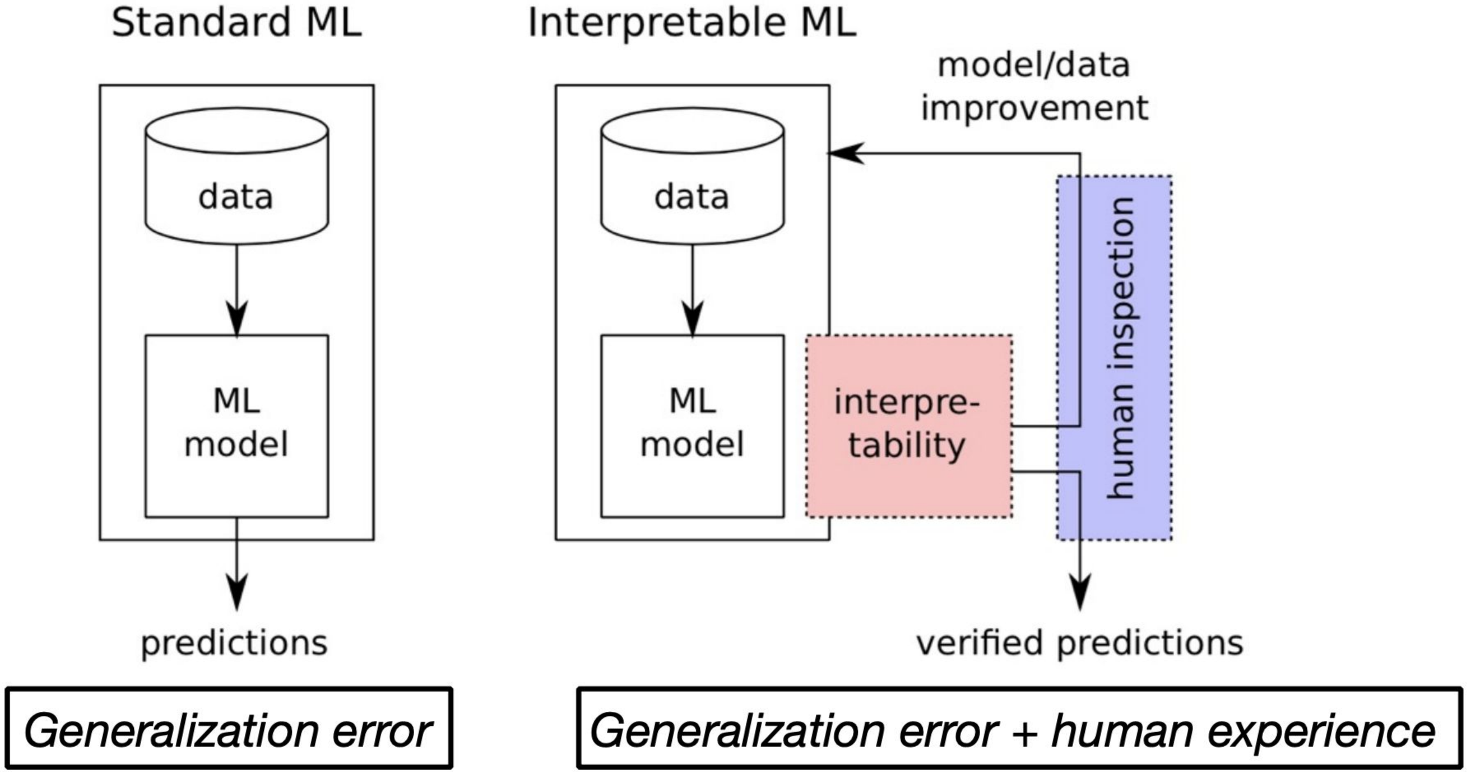
Explanation - From a Model Perspective

Why Explainability: Debug (Mis-)Predictions

Top label: **"clog"**

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

Why Explainability: Improve ML Model



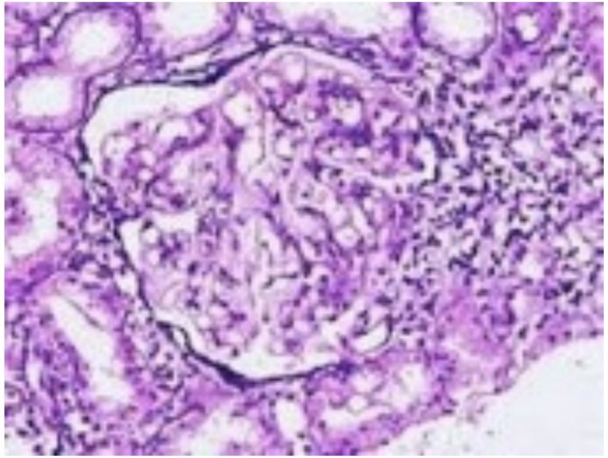
Credit: Samek, Binder, Tutorial on Interpretable ML, MICCAI'18

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 …"

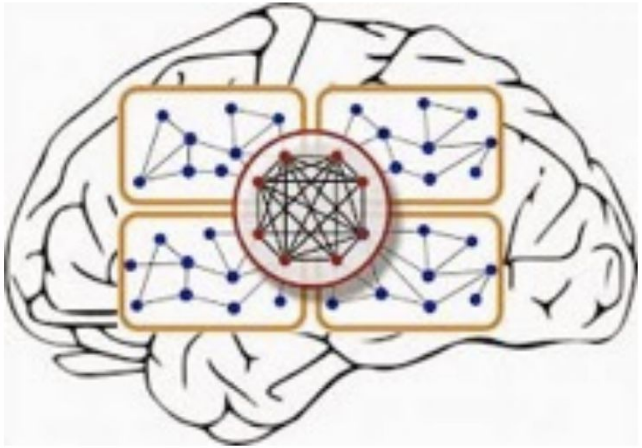


Credit: Samek, Binder, Tutorial on Interpretable ML, MICCAI'18

Why Explainability: Learn New Insights

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

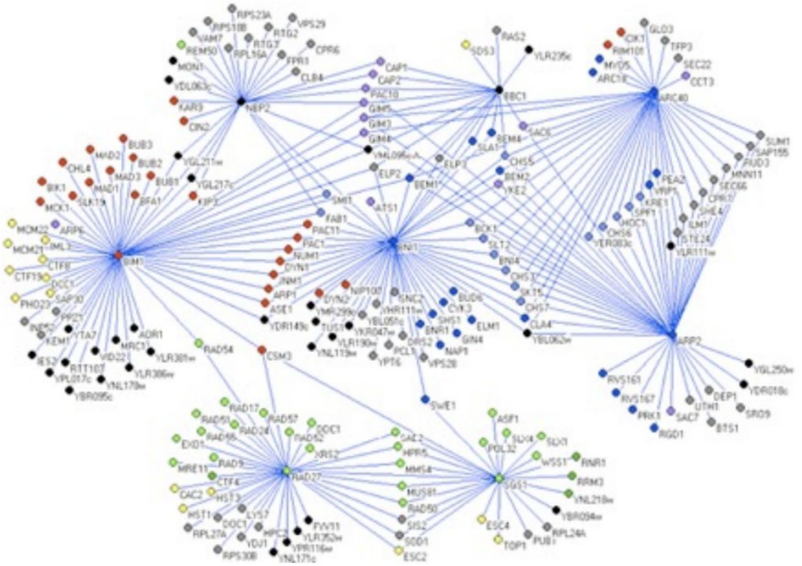
Old promise:
“Learn about the human brain.”

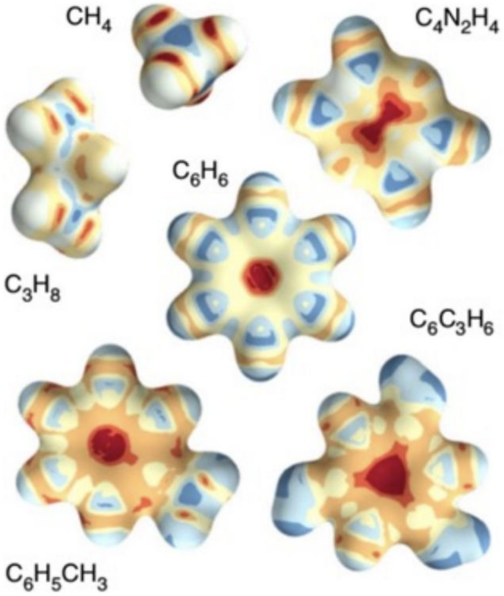


Credit: Samek, Binder, Tutorial on Interpretable ML, MICCAI’18

Why Explainability: Learn Insights in the Sciences

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





Credit: Samek, Binder, Tutorial on Interpretable ML, MICCAI'18

Explanation - From a Regulatory Perspective

Why Explainability: Laws against Discrimination

Citizenship

Immigration Reform and Control Act

Sex

Equal Pay Act of 1963;
Civil Rights Act of 1964

Age

Age Discrimination in Employment Act of 1967

Race

Civil Rights Act of 1964

Disability status

Rehabilitation Act of 1973;
Americans with Disabilities Act of 1990

And more...

SR 11-7: Guidance on Model Risk Management

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

Fairness

Privacy

GDPR

Transparency

CALIFORNIA
CONSUMER
PRIVACY
ACT OF 2018

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 Commision

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 Commision

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.

Fai

cy

Recital 71

Profiling*

[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,

Transparenc y

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 zip code.

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**.

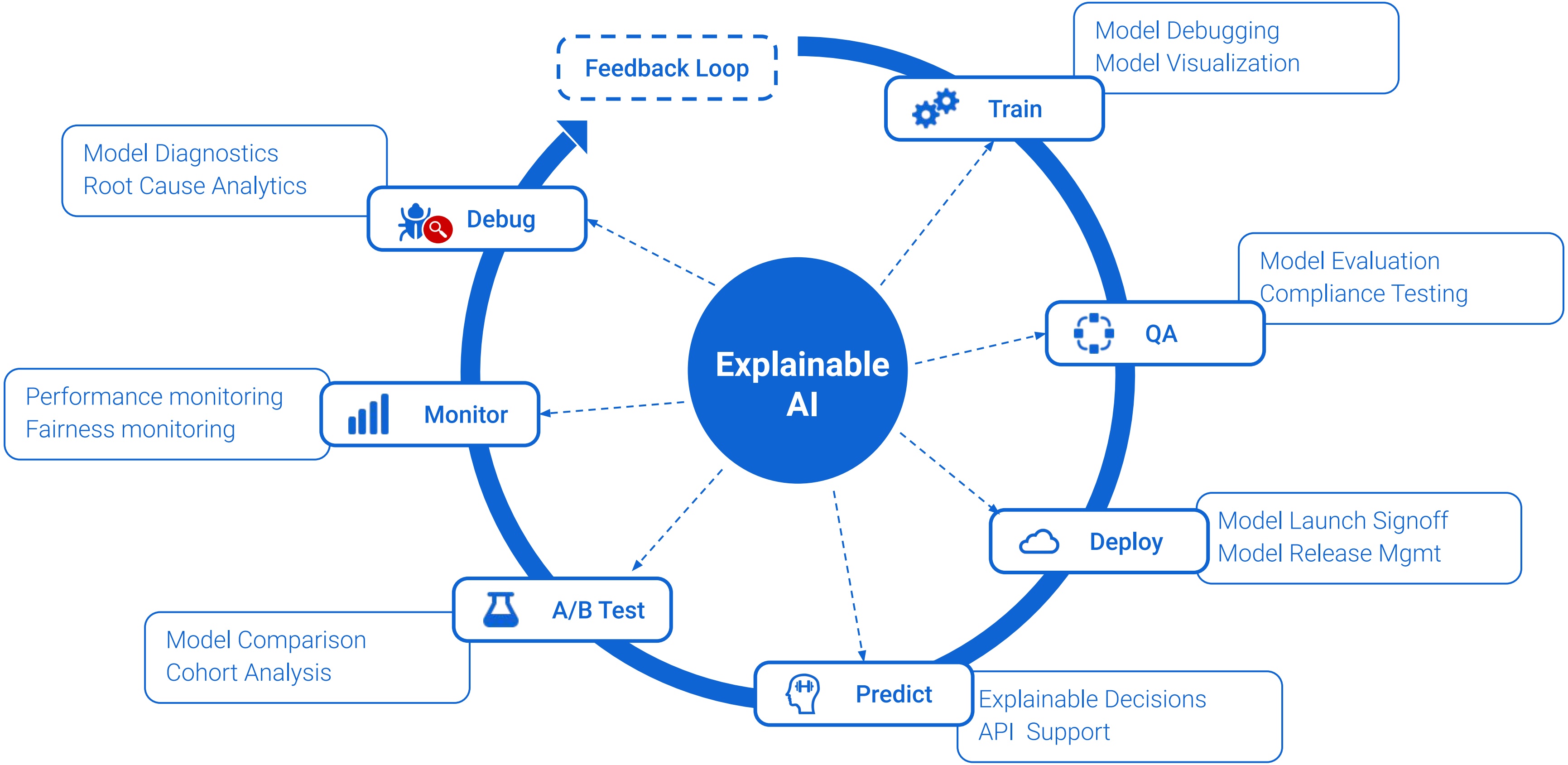
In addition **SR12-07** calls for **incorporating validation or other type of independent review of the stress testing framework to ensure the integrity of stress testing processes and results.**

JOHN HILL

GLOBAL HEAD OF MODEL RISK GOVERNANCE, CREDIT SUISSE

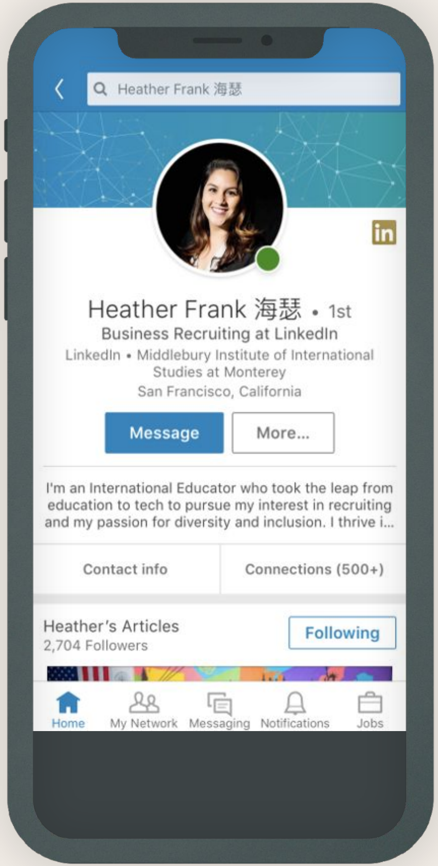
" In the current regulatory environment, model validation policies must be fully compliant with the requirements of SR11-7. While SR11-7 officially applies to US conforming bank and non-US banks doing business in the US, many European financial firms have adopted SR11-7 as their standard as well. "

“Explainability by Design” for AI products

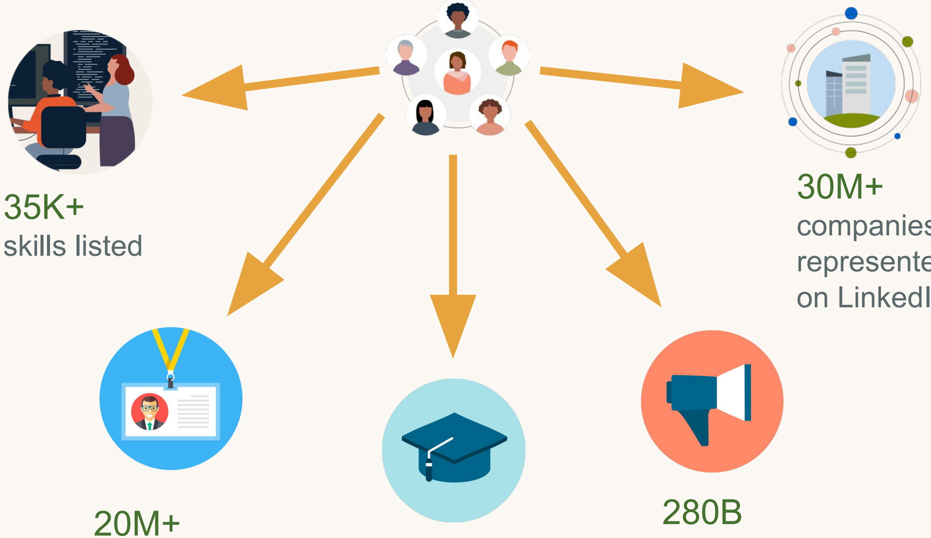


AI @ Scale - Challenges for Explainable AI

LinkedIn operates the largest professional network on the Internet



645M+ members



35K+
skills listed

30M+
companies are represented on LinkedIn

20M+
open jobs on LinkedIn Jobs

90K+
schools listed (high school & college)

280B
Feed updates

The AWS ML Stack

Broadest and most complete set of Machine Learning capabilities

AI SERVICES

VISION	SPEECH		TEXT			SEARCH NEW	CHATBOTS	PERSONALIZATION	FORECASTING	FRAUD NEW	DEVELOPMENT NEW	CONTACT CENTERS NEW
Amazon Rekognition	Amazon Polly	Amazon Transcribe NEW +Medical	Amazon Comprehend +Medical	Amazon Translate	Amazon Textract	Amazon Kendra	Amazon Lex	Amazon Personalize	Amazon Forecast	Amazon Fraud Detector	Amazon CodeGuru	Contact Lens *For Amazon Connect*

ML SERVICES

Amazon SageMaker

Ground Truth | Augmented AI

SageMaker Studio IDE NEW

Built-in algorithms	SageMaker Notebooks NEW	SageMaker Experiments NEW	Model tuning	SageMaker Debugger NEW	SageMaker Autopilot NEW	Model hosting	SageMaker Model Monitor NEW

SageMaker Neo

ML FRAMEWORKS & INFRASTRUCTURE

TensorFlow NEW | mxnet NEW | PYTORCH NEW

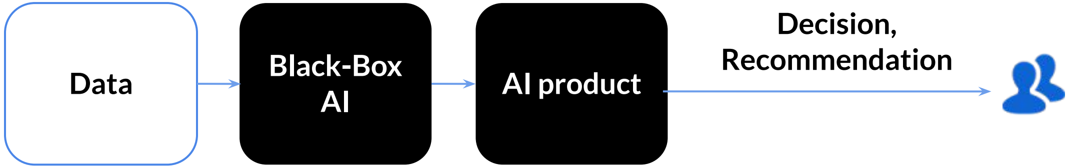
GLUON | Keras | HOROVOD

Deep Learning AMIs & Containers | GPUs & CPUs | Elastic Inference | Inferentia | FPGA

Explanation - In a Nutshell

What is Explainable AI?

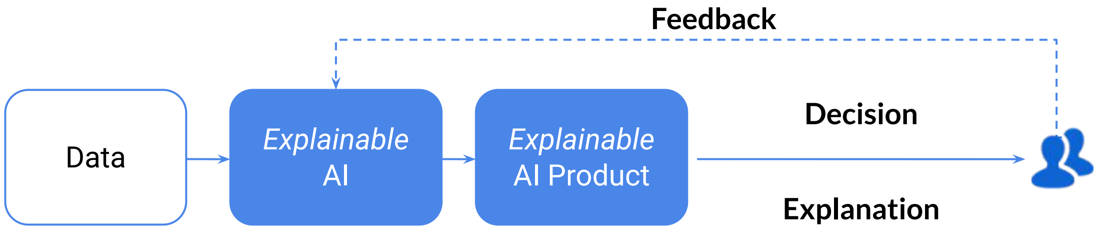
Black Box AI



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?

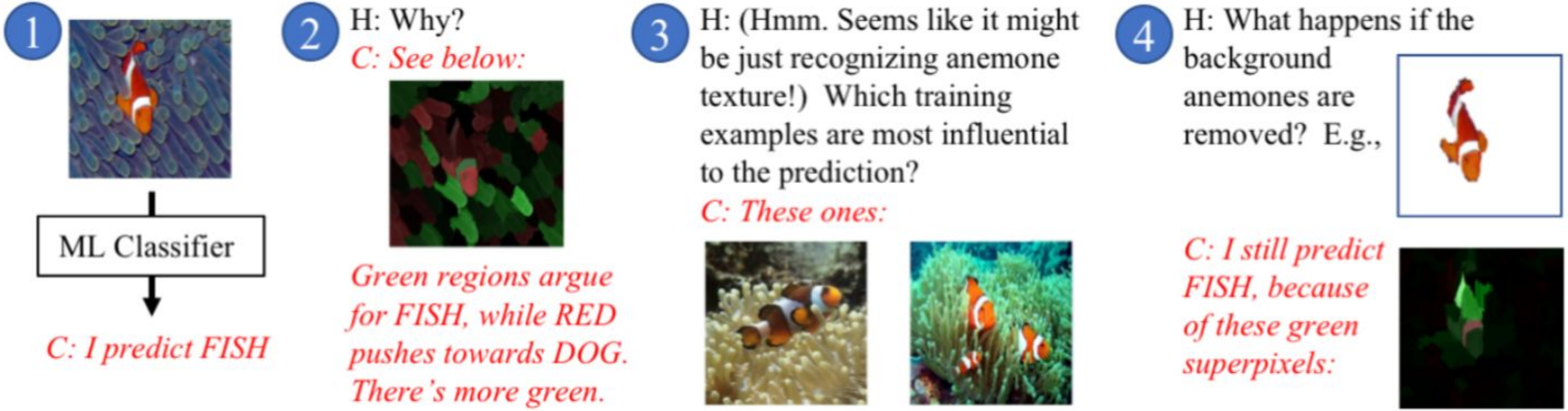
Explainable AI



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



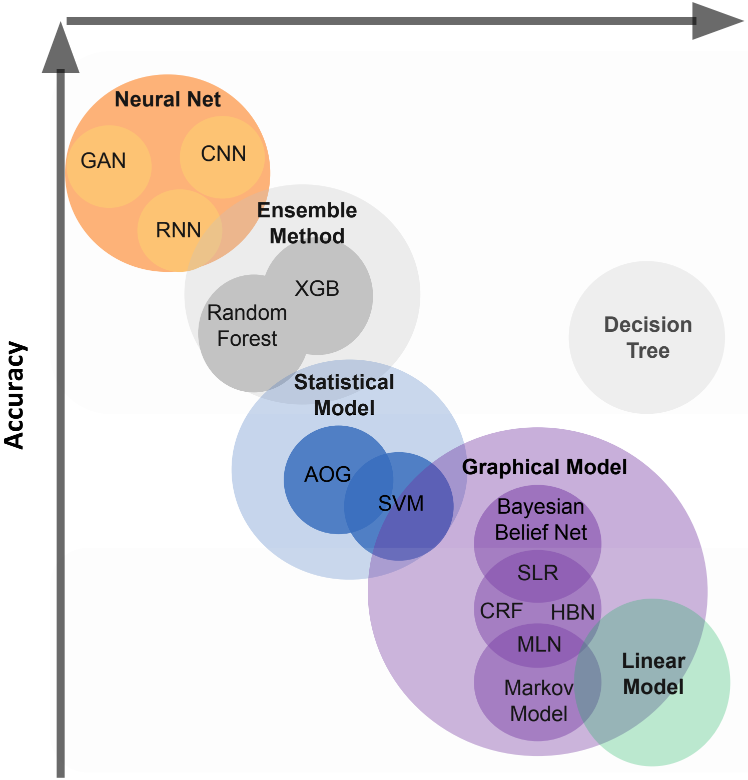
- Humans may have follow-up questions
- Explanations cannot answer all users' concerns

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

How to Explain? Accuracy vs. Explainability

Learning

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



Interpretability

Non-Linear functions

Polynomial functions

Quasi-Linear functions

XAI Definitions - Explanation vs. Interpretation

Oxford Dictionary of English

explanation | ɛksplə'neɪʃ(ə)n |

noun

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

interpret | ɪn'təːprɪt |

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

Field names

One row (4 fields)

2000 rows all told

Tabular

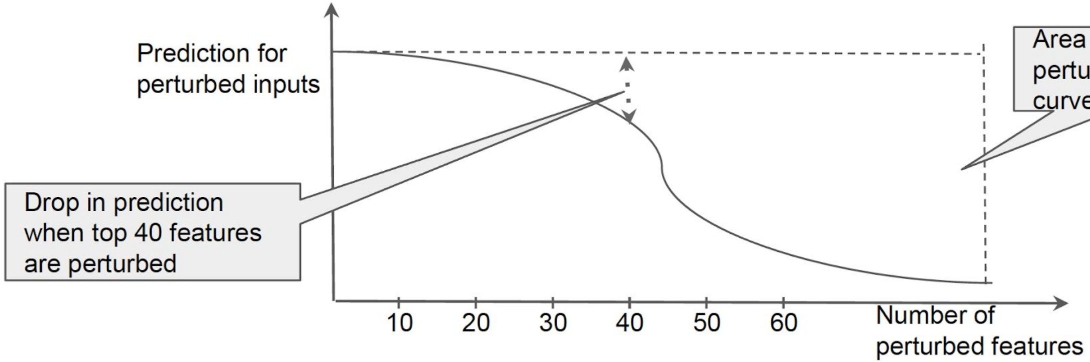
Images

Text

Evaluation (1) - Perturbation-based Approaches

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

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



Evaluation (2) - Human (Role)-based Evaluation is Essential… but too often based on size!

Evaluation criteria for Explanations [Miller, 2017]

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

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

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

[Doshi-Velez and Kim 2017, Poursabzi-Sangdeh 18]

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

Comprehensibility

How much effort for correct human interpretation?

Succinctness

How concise and compact is the explanation?

Actionability

What can one action, do with the explanation?

Reusability

Could the explanation be personalized?

Accuracy

How accurate and precise is the explanation?

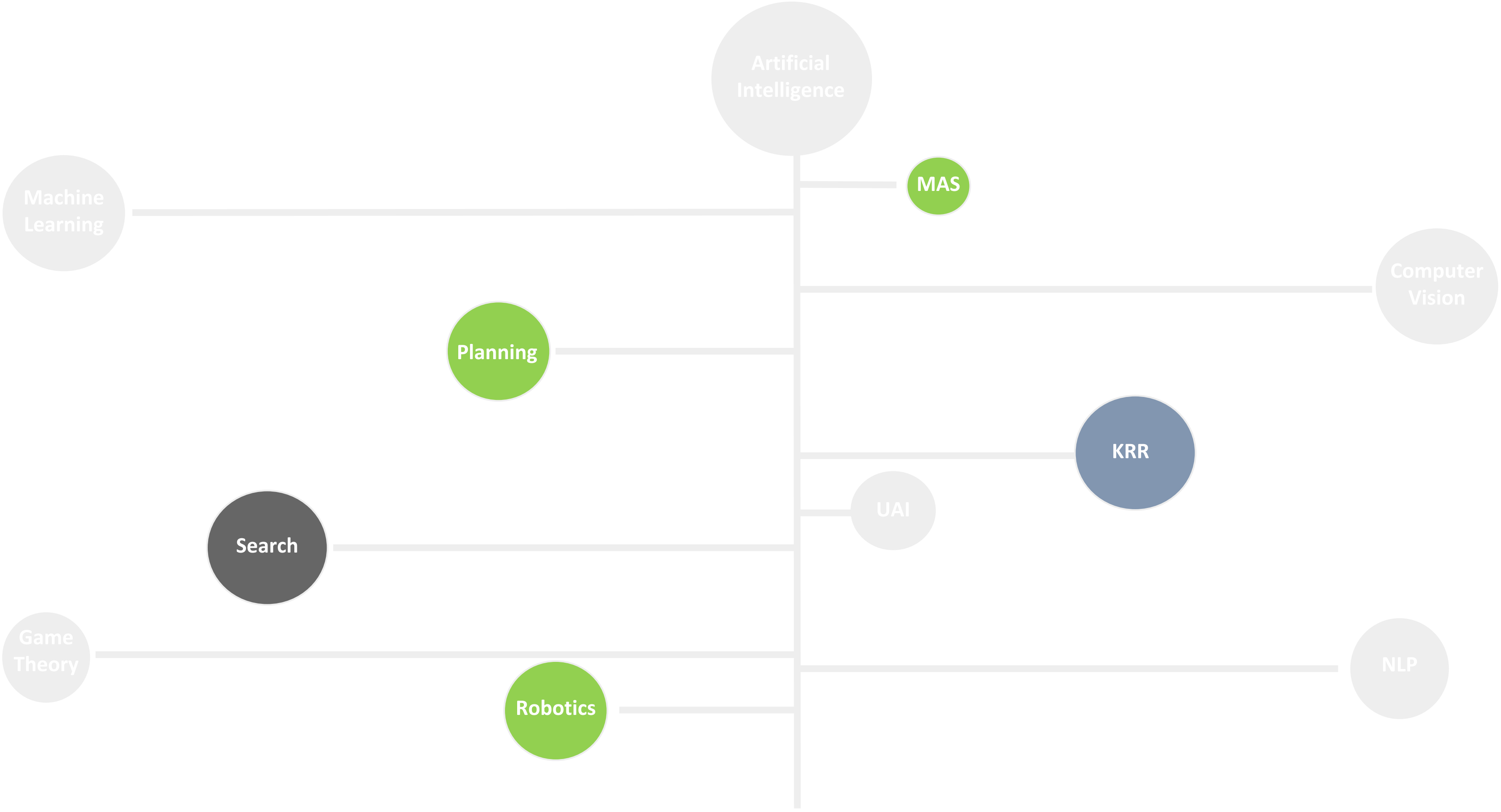
Completeness

Is the explanation complete, partial, restricted?

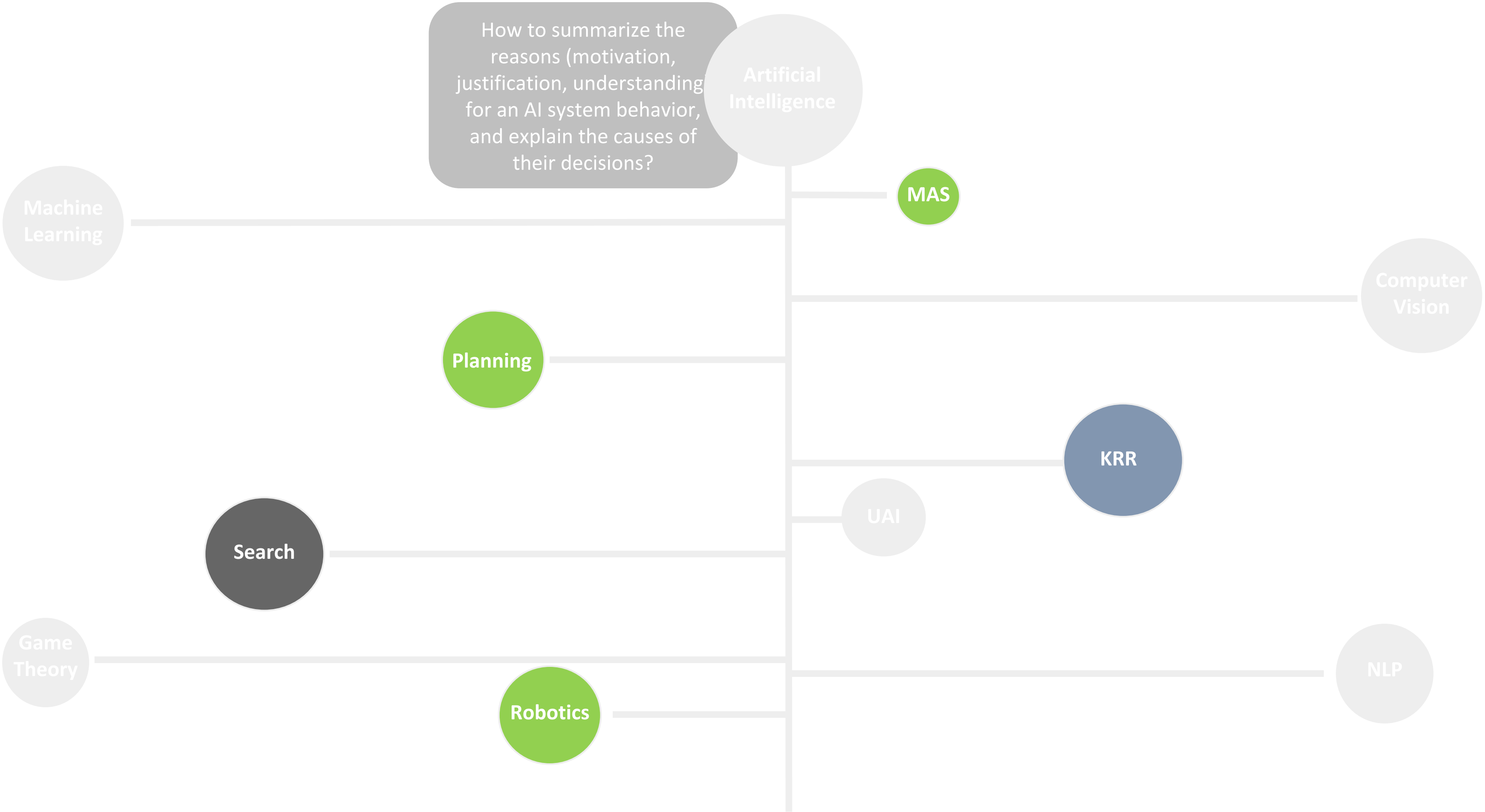
Source: Accenture Point of View. Understanding Machines: Explainable AI. Freddy Lecue, Dadong Wan

Explanation in AI (not only Machine Learning!)

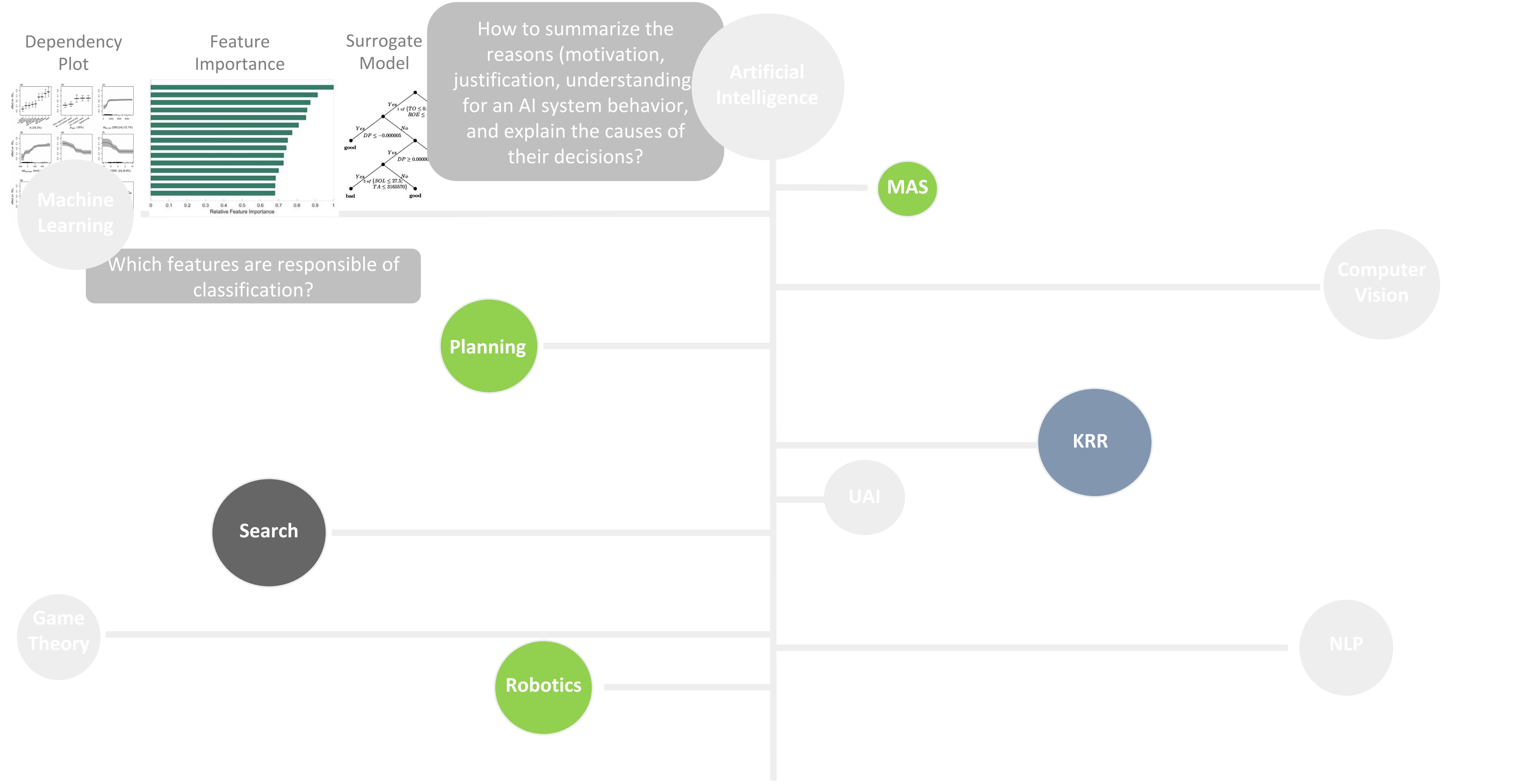
XAI: One Objective, Many ‘AI’s, Many Definitions, Many Approaches



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

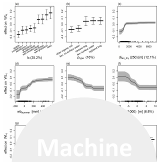


XAI: One Objective, Many 'AI's, Many Definitions, Many Approaches

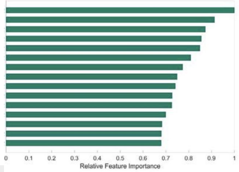


XAI: One Objective, Many 'AI's, Many Definitions, Many Approaches

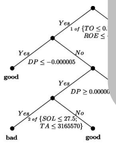
Dependency Plot



Feature Importance



Surrogate Model



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

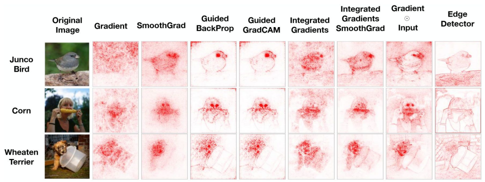
Artificial Intelligence

Machine Learning

Which features are responsible of classification?

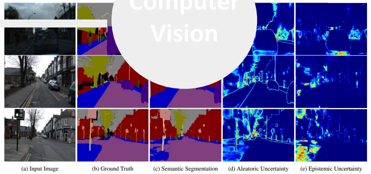
MAS

Saliency Map



Which complex features are responsible of classification?

Computer Vision



Uncertainty Map

Planning

KRR

UAI

Search

Game Theory

Robotics

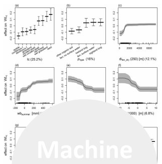
NLP

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

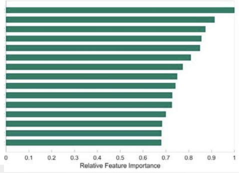
How to summarize the reasons (motivation, justification, understanding for an AI system behavior, and explain the causes of their decisions?

Artificial Intelligence

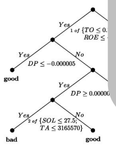
Dependency Plot



Feature Importance



Surrogate Model



Machine Learning

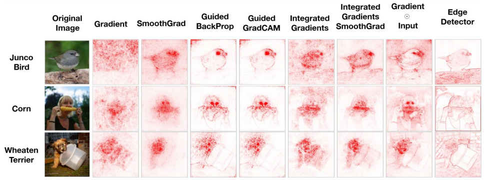
Which features are responsible of classification?

Strategy Summarization

MAS

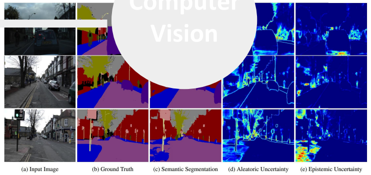
- Which agent strategy & plan ?
- Which player contributes most?
- Why such a conversational flow?

Saliency Map



Which complex features are responsible of classification?

Computer Vision



Uncertainty Map

Planning

KRR

UAI

Search

Game Theory

Robotics

NLP

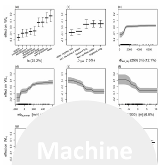
XAI: One Objective, Many 'AI's, Many Definitions, Many Approaches

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

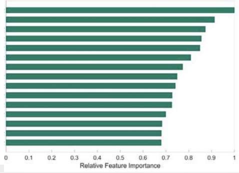
Artificial Intelligence

Machine Learning

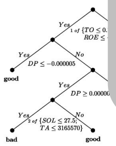
Dependency Plot



Feature Importance



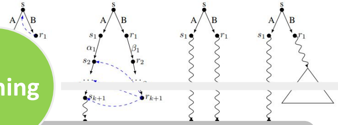
Surrogate Model



Which features are responsible of classification?

Planning

Plan Refinement



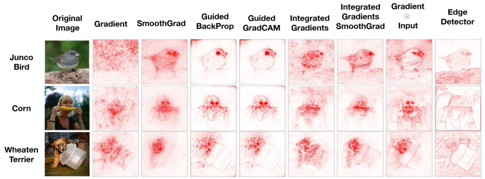
Which actions are responsible of a plan?

MAS

Strategy Summarization

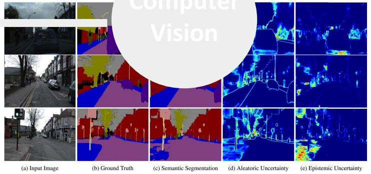
- Which agent strategy & plan ?
- Which player contributes most?
- Why such a conversational flow?

Saliency Map



Which complex features are responsible of classification?

Computer Vision



Uncertainty Map

KRR

UAI

Search

Game Theory

Robotics

NLP

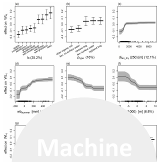
XAI: One Objective, Many 'AI's, Many Definitions, Many Approaches

Artificial Intelligence

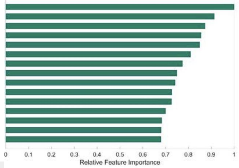
How to summarize the reasons (motivation, justification, understanding for an AI system behavior, and explain the causes of their decisions?

Machine Learning

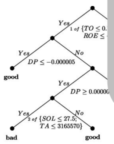
Dependency Plot



Feature Importance



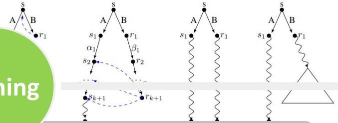
Surrogate Model



Which features are responsible of classification?

Planning

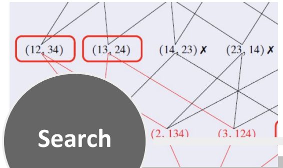
Plan Refinement



Which actions are responsible of a plan?

Search

Conflicts Resolution



Which constraints can be relaxed?

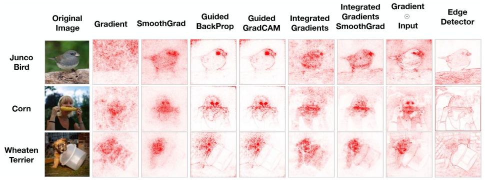
MAS

Strategy Summarization

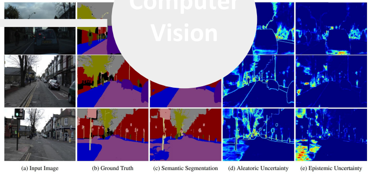
- Which agent strategy & plan ?
- Which player contributes most?
- Why such a conversational flow?

Computer Vision

Saliency Map



Which complex features are responsible of classification?



Uncertainty Map

KRR

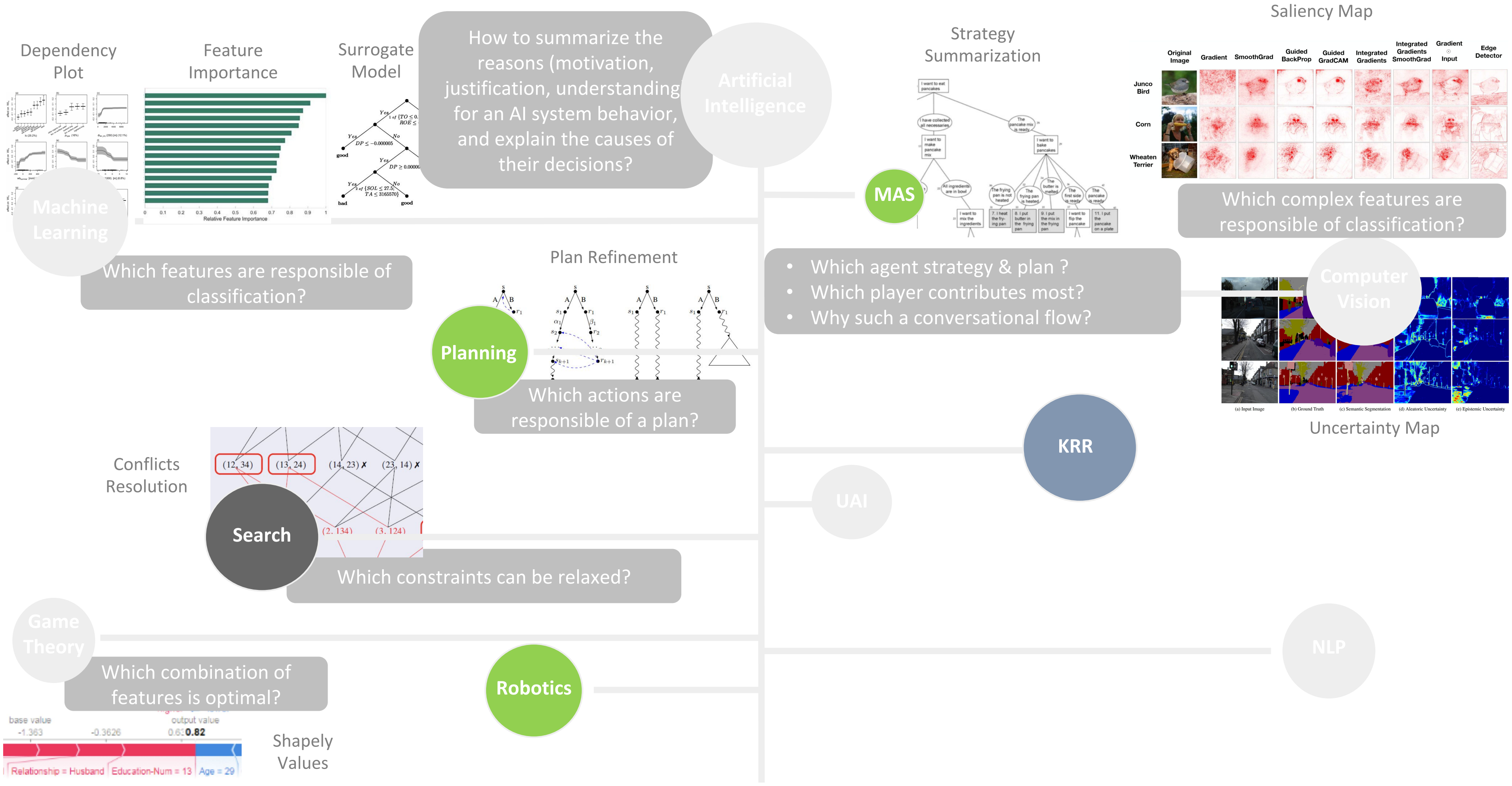
UAI

Game Theory

Robotics

NLP

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



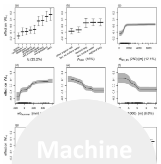
XAI: One Objective, Many 'AI's, Many Definitions, Many Approaches

Artificial Intelligence

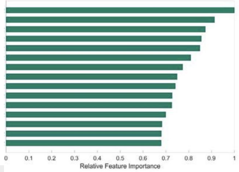
How to summarize the reasons (motivation, justification, understanding for an AI system behavior, and explain the causes of their decisions?

Machine Learning

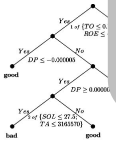
Dependency Plot



Feature Importance



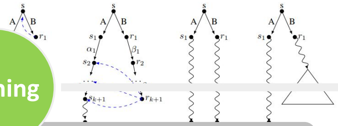
Surrogate Model



Which features are responsible of classification?

Planning

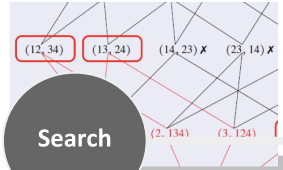
Plan Refinement



Which actions are responsible of a plan?

Search

Conflicts Resolution



Which constraints can be relaxed?

Game Theory

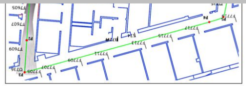
Which combination of features is optimal?

Shapely Values

Robotics

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

Narrative-based



MAS

Strategy Summarization

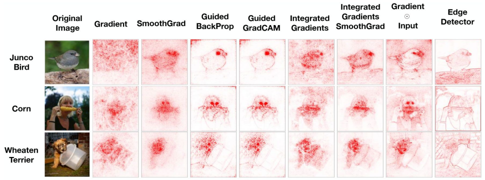
- Which agent strategy & plan ?
- Which player contributes most?
- Why such a conversational flow?

KRR

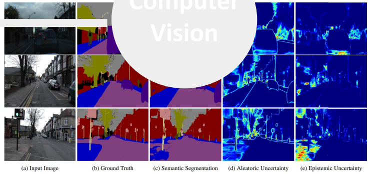
UAI

Computer Vision

Saliency Map



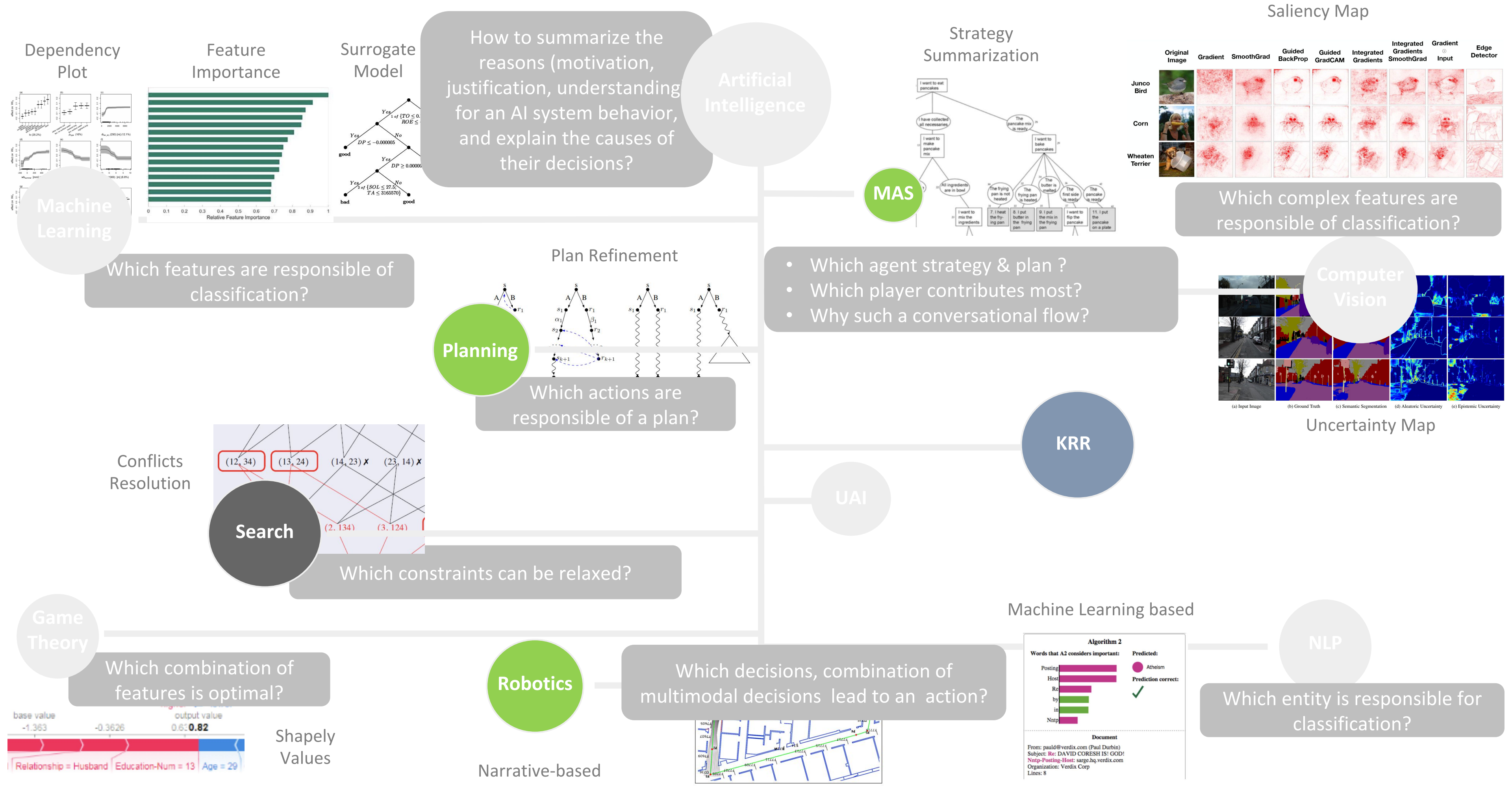
Which complex features are responsible of classification?



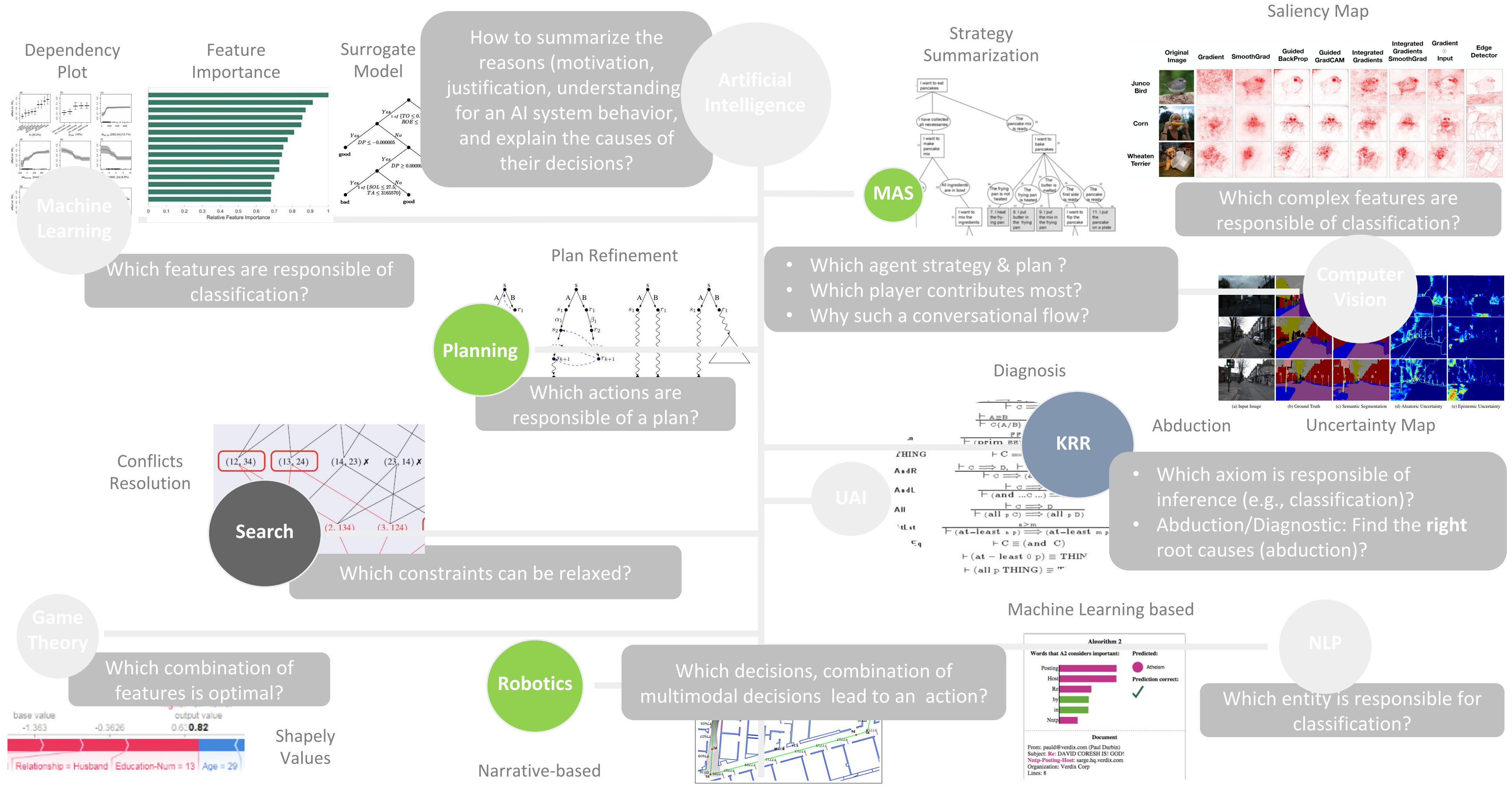
Uncertainty Map

NLP

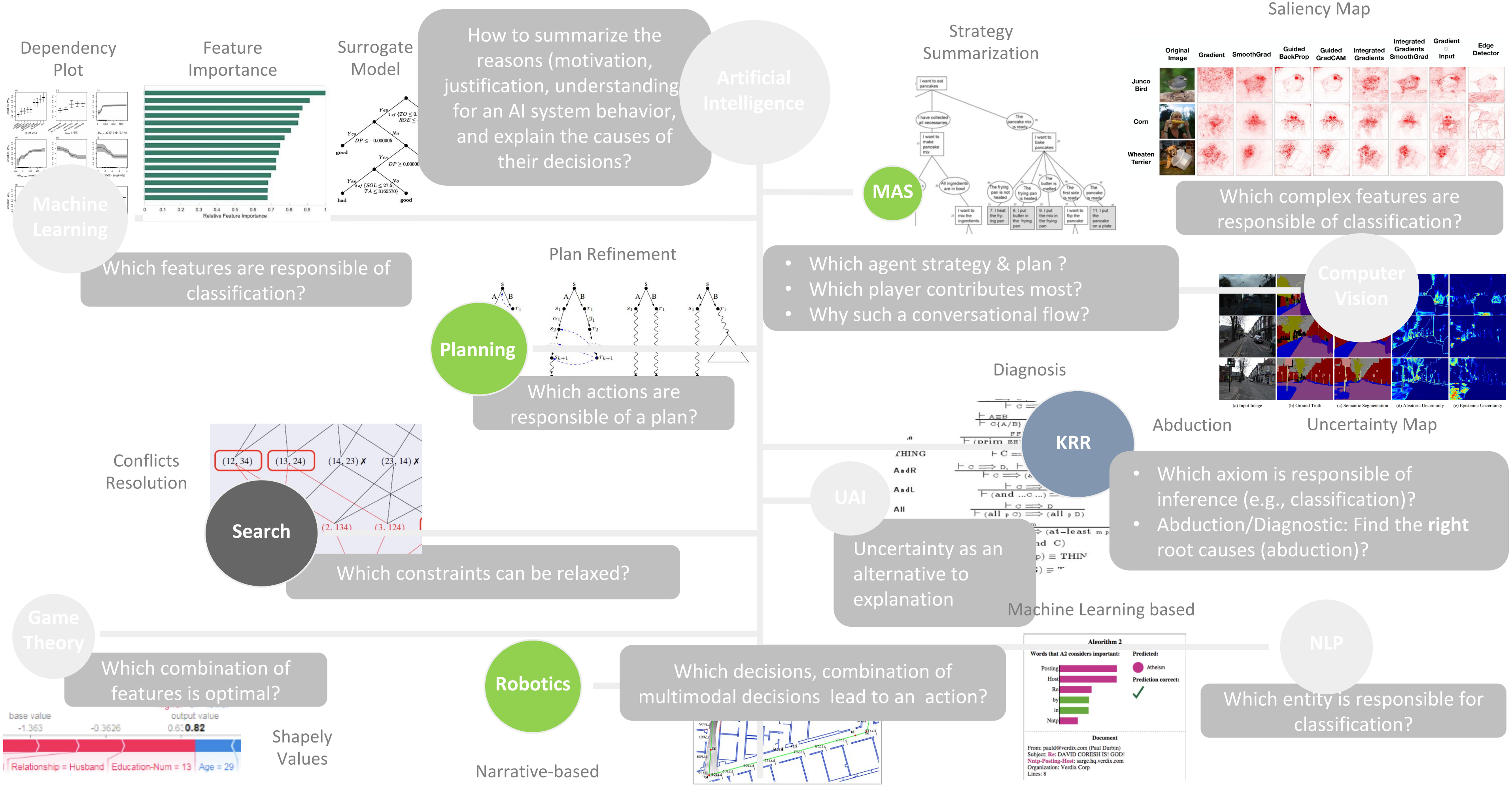
XAI: One Objective, Many ‘AI’s, Many Definitions, Many Approaches



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



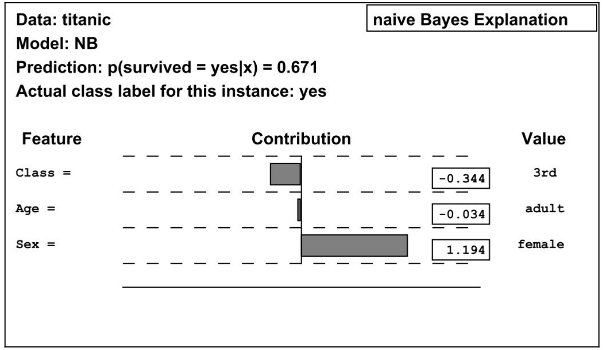
XAI: One Objective, Many 'AI's, Many Definitions, Many Approaches



Overview of Explanation in Machine Learning (1)

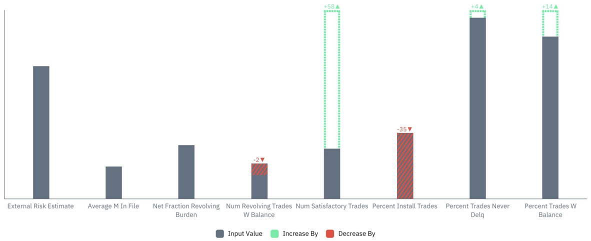
Interpretable Models:

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



Naive Bayes model

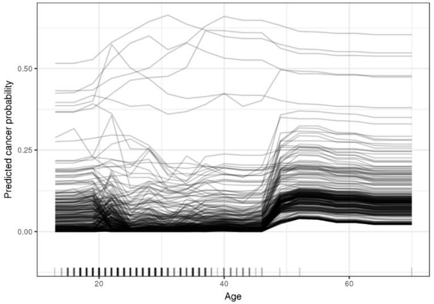
Igor Kononenko. Machine learning for medical diagnosis: history, state of the art and perspective. Artificial Intelligence in Medicine, 23:89–109, 2001.

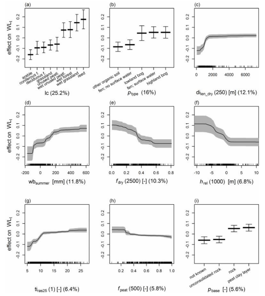


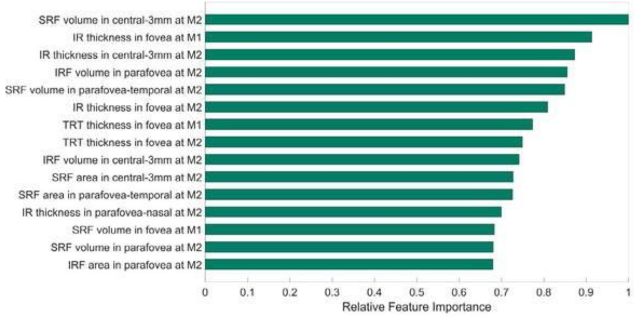
Counterfactual What-if

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

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





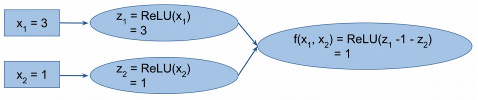


(a)

Feature Importance
Partial Dependence Plot
Individual Conditional Expectation
Sensitivity Analysis

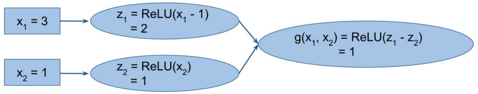
Overview of Explanation in Machine Learning (2)

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



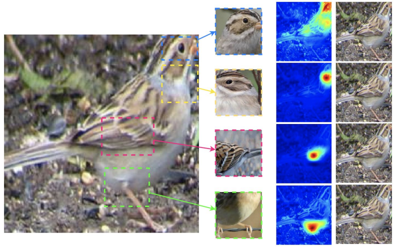
Network $g(x_1, x_2)$
Attributions at $x_1 = 3, x_2 = 1$

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

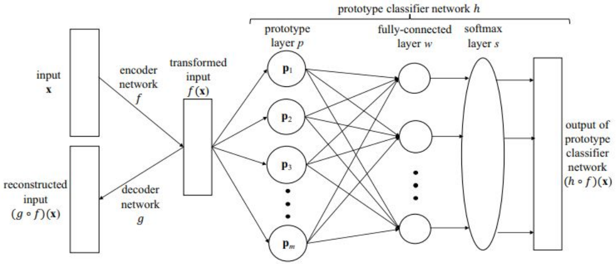
Attribution for Deep Network (Integrated gradient-based)

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

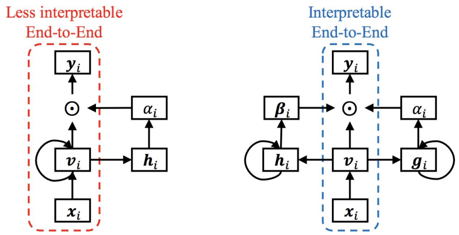


Chaofan Chen, Oscar Li, Alina Barnett, Jonathan Su, Cynthia Rudin: This looks like that: deep learning for interpretable image recognition. CoRR abs/1806.10574 (2018)



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

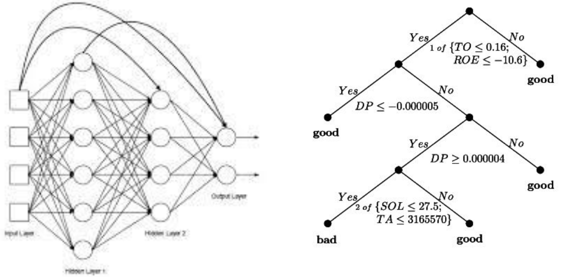


(a) Standard attention model (b) RETAIN model

Attention Mechanism

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

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

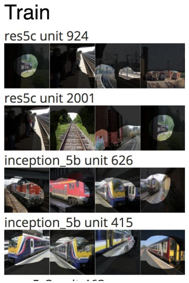


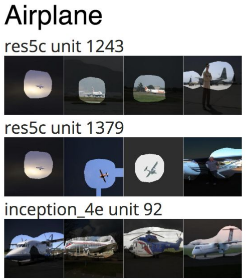
Surogate Model

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

Overview of Explanation in Machine Learning (3)

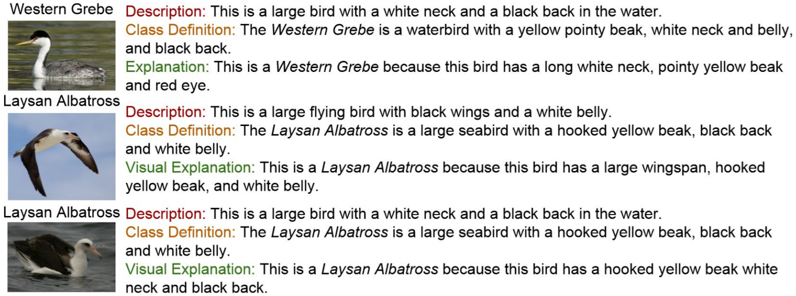
- Computer Vision





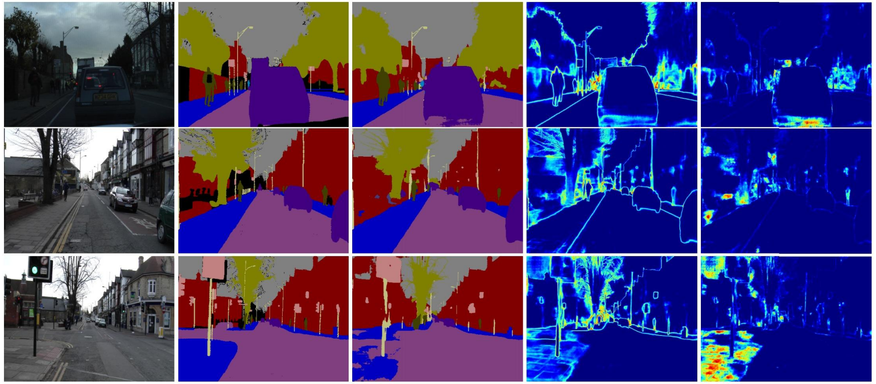
Interpretable Units

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



Visual Explanation

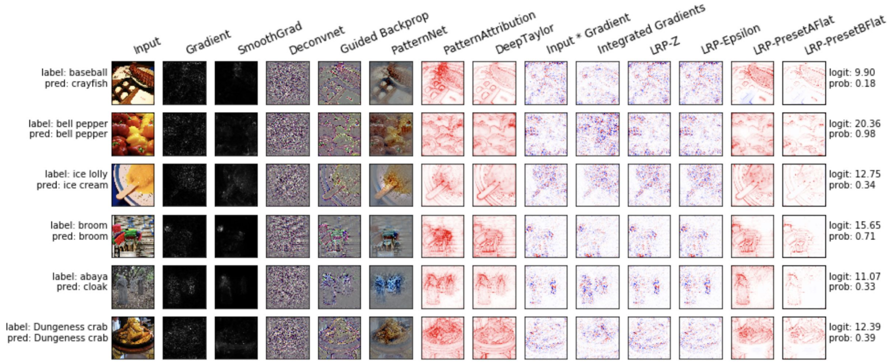
Lisa Anne Hendricks, Zeynep Akata, Marcus Rohrbach, Jeff Donahue, Bernt Schiele, Trevor Darrell: Generating Visual Explanations. ECCV (4) 2016: 3-19



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

Uncertainty Map

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

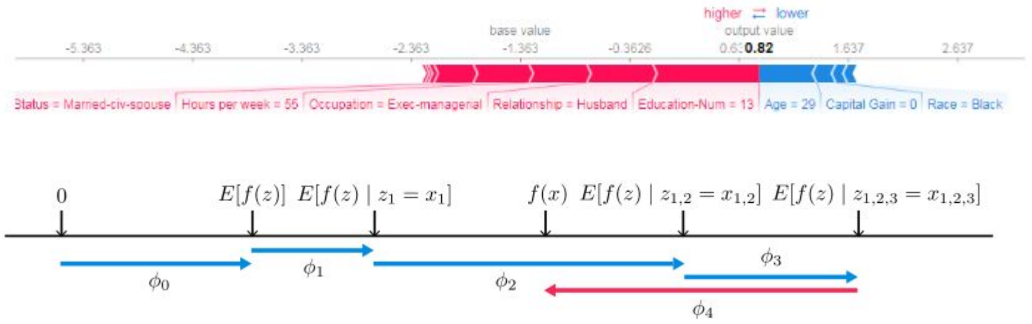


Saliency Map

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

Overview of Explanation in Different AI Fields (1)

- Game Theory

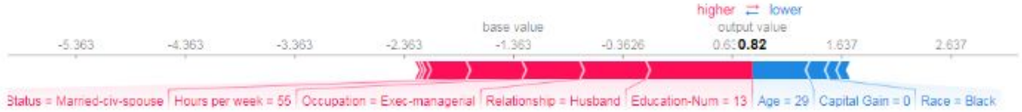


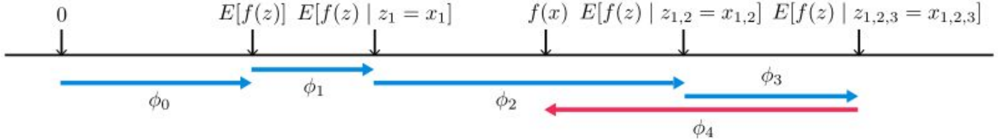
Shapley Additive Explanation

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

Overview of Explanation in Different AI Fields (1)

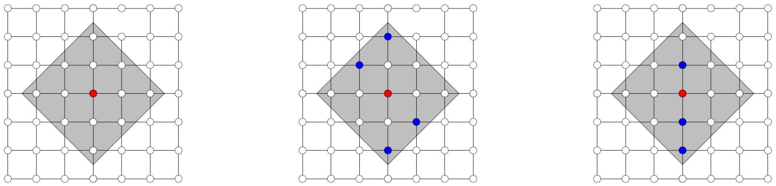
- Game Theory





Shapley Additive Explanation

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

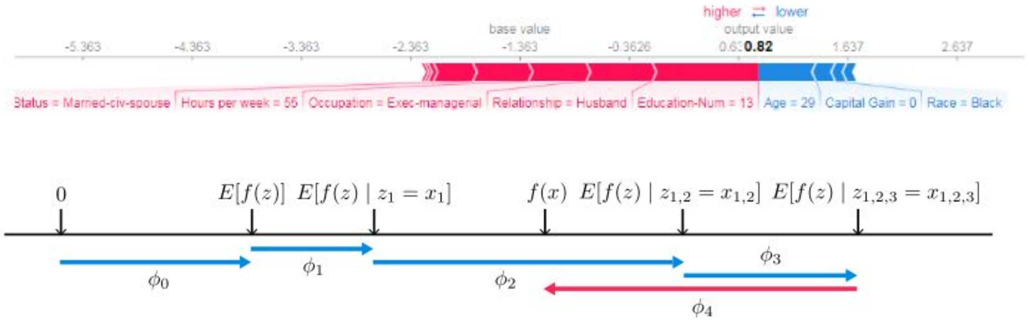


L-Shapley and C-Shapley (with graph structure)

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

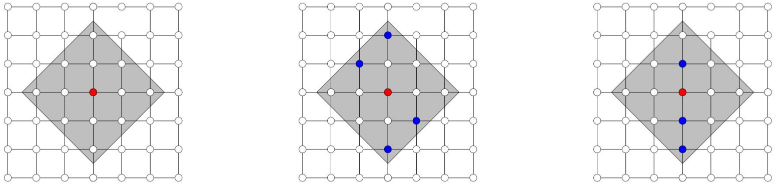
Overview of Explanation in Different AI Fields (1)

- Game Theory



Shapley Additive Explanation

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



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

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

instance-wise feature importance (causal influence)

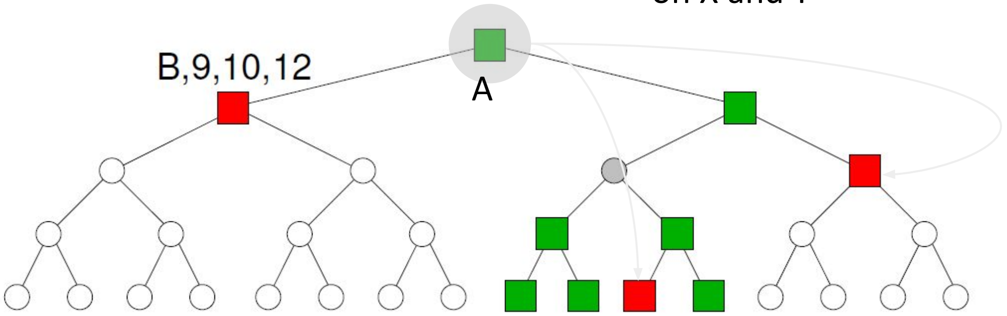
Erik Štrumbelj and Igor Kononenko. An efficient explanation of individual classifications using game theory. Journal of Machine Learning Research, 11:1–18, 2010.

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

If A+1 then NEW Conflicts
on X and Y



Conflicts resolution

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

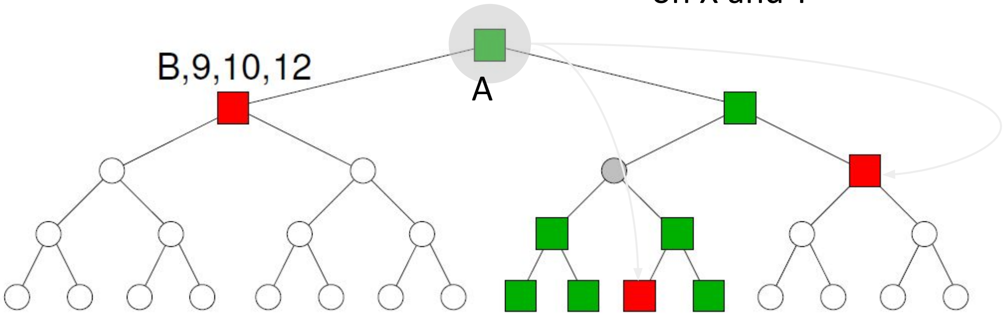
Robustness Computation

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

Overview of Explanation in Different AI Fields (2)

- Search and Constraint Satisfaction

If A+1 then NEW Conflicts on X and Y

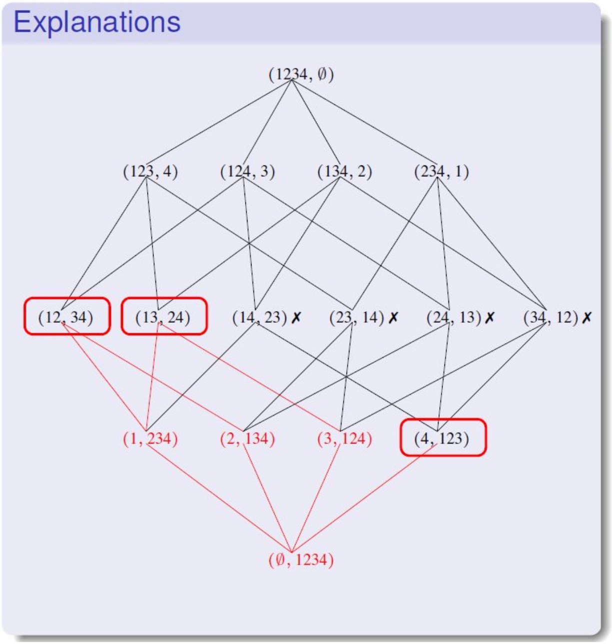


Conflicts resolution

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

Robustness Computation

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



Constraints relaxation

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

Overview of Explanation in Different AI Fields (3)

- Knowledge Representation and Reasoning

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

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

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

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

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

Overview of Explanation in Different AI Fields (3)

- Knowledge Representation and Reasoning

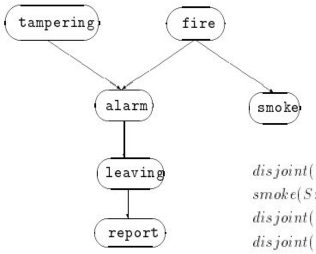
Ref	$\vdash C \Longrightarrow C$
Trans	$\dfrac{\vdash C \Longrightarrow D,\ \vdash D \Longrightarrow E}{\vdash C \Longrightarrow E}$
Eq	$\dfrac{\vdash A \equiv B \qquad \vdash C \Longrightarrow D}{\vdash C\{A/B\} \Longrightarrow D\{A/B\}}$
Prim	$\dfrac{FF \subset EE}{\vdash (\text{prim } EE) \Longrightarrow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

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

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

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

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



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

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

Abduction Reasoning (in Bayesian Network)

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

Overview of Explanation in Different AI Fields (3)

- Knowledge Representation and Reasoning

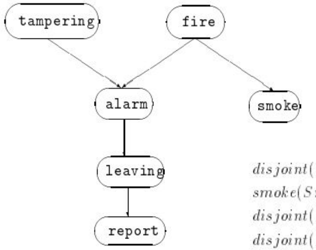
Ref	$\vdash C \Longrightarrow C$
Trans	$\dfrac{\vdash C \Longrightarrow D,\ \vdash D \Longrightarrow E}{\vdash C \Longrightarrow E}$
Eq	$\dfrac{\vdash A \equiv B \qquad \vdash C \Longrightarrow D}{\vdash C\{A/B\} \Longrightarrow D\{A/B\}}$
Prim	$\dfrac{FF \subset EE}{\vdash (\text{prim } EE) \Longrightarrow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

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

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

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

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



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

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

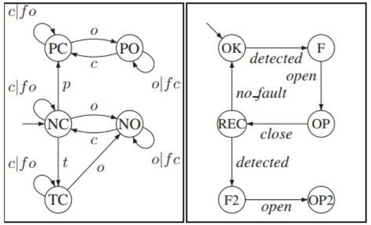
$smoke(Sm) \leftarrow fire(Fi) \wedge c\_smoke(Sm, Fi).$

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

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

Abduction Reasoning (in Bayesian Network)

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

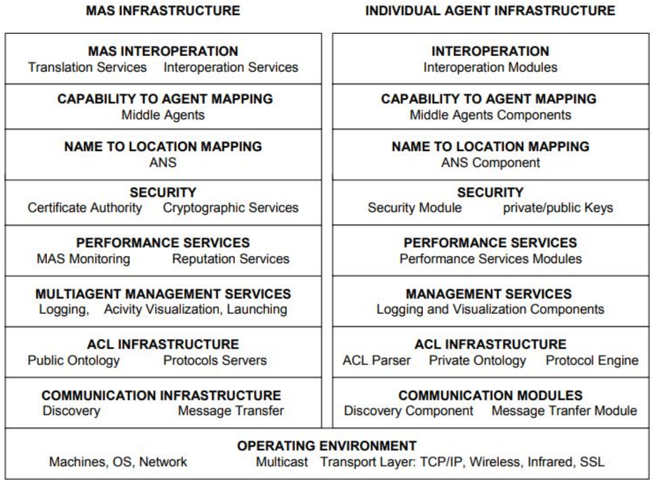


Diagnosis Inference

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

Overview of Explanation in Different AI Fields (4)

- Multi-agent Systems

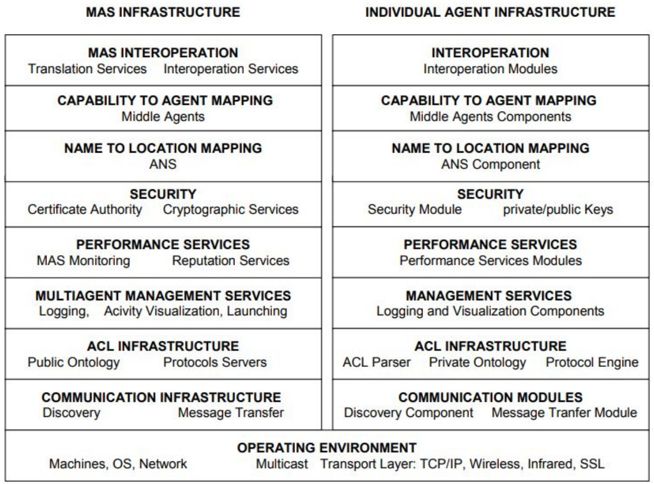


Explanation of Agent Conflicts & Harmful Interactions

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

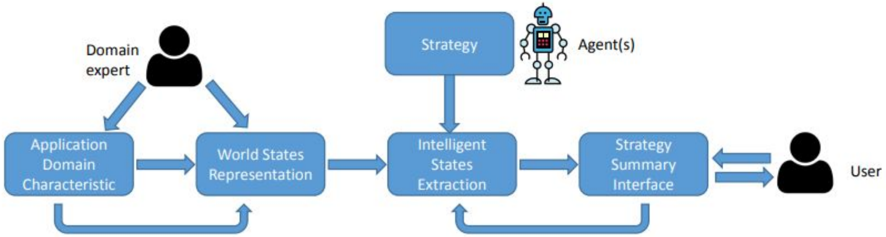
Overview of Explanation in Different AI Fields (4)

- Multi-agent Systems



Explanation of Agent Conflicts & Harmful Interactions

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

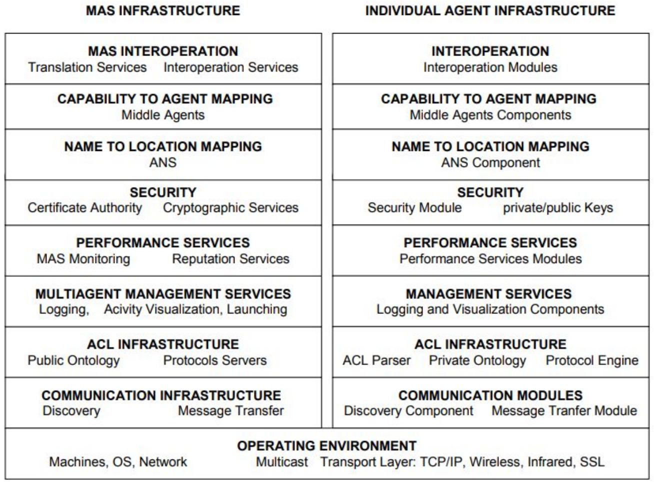


Agent Strategy Summarization

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

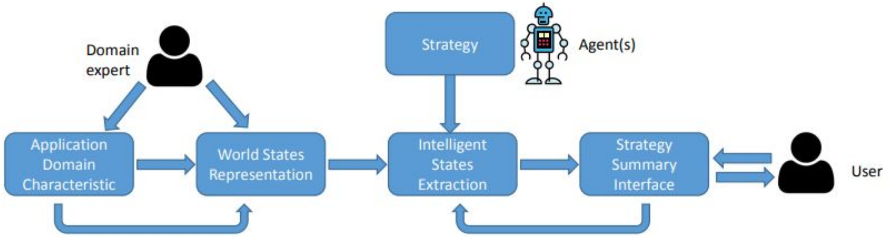
Overview of Explanation in Different AI Fields (4)

- Multi-agent Systems



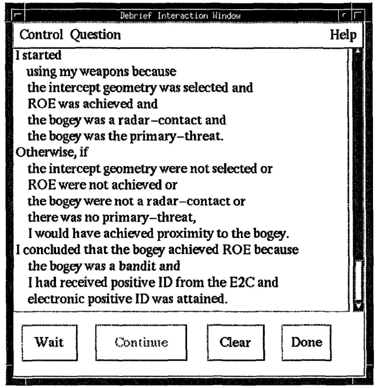
Explanation of Agent Conflicts & Harmful Interactions

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



Agent Strategy Summarization

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



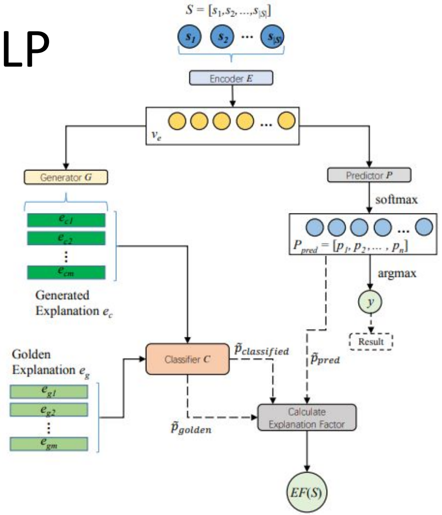
Explainable Agents

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



Explainable NLP

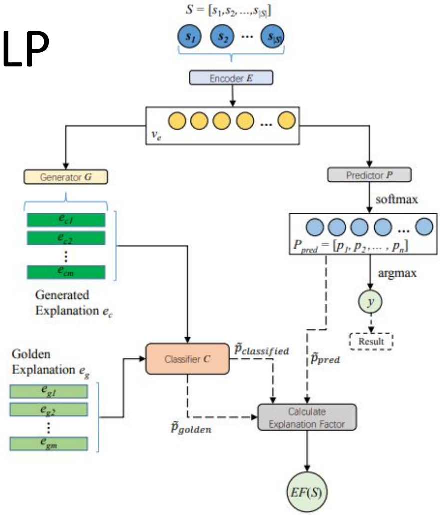
Hui Liu, Qingyu Yin, William Yang Wang: Towards Explainable NLP: A Generative Explanation Framework for Text Classification. CoRR abs/1811.00196 (2018)

Fine-grained explanations are in the form of:

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

Overview of Explanation in Different AI Fields (5)

- NLP

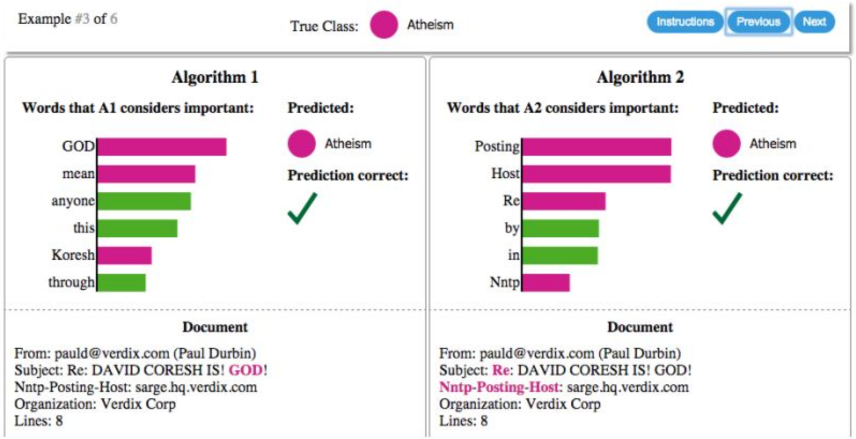


Fine-grained explanations are in the form of:

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

Explainable NLP

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

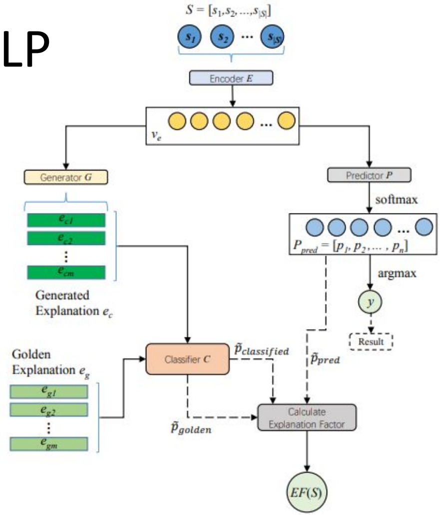


LIME for NLP

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

Overview of Explanation in Different AI Fields (5)

- NLP

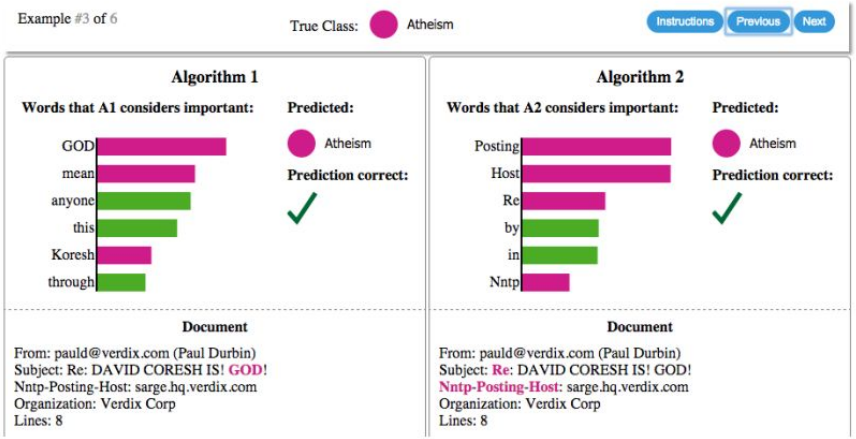


Explainable NLP

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

Fine-grained explanations are in the form of:

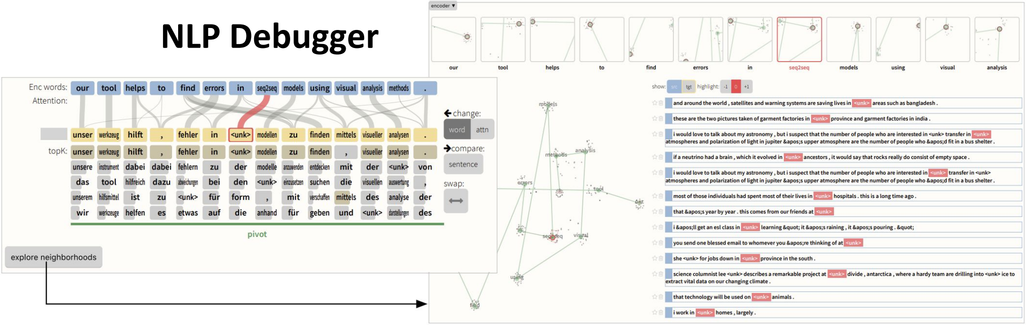
- texts in a real-world dataset;
- Numerical scores



LIME for NLP

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

NLP Debugger



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

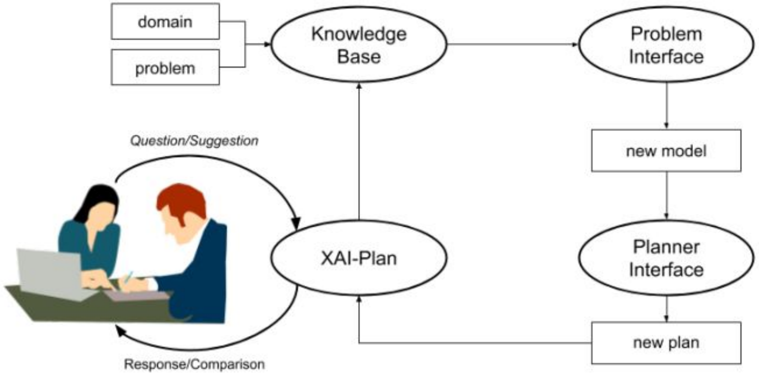
Hendrik Strobelt, Sebastian Gehrmann, Michael Behrisch, Adam Perer, Hanspeter Pfister, Alexander M. Rush: Seq2seq-Vis: A Visual Debugging Tool for Sequence-to-Sequence Models. IEEE Trans. Vis. Comput. Graph. 25(1): 353-363 (2019)

Overview of Explanation in Different AI Fields (6)

- Planning and Scheduling

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

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



XAI Plan

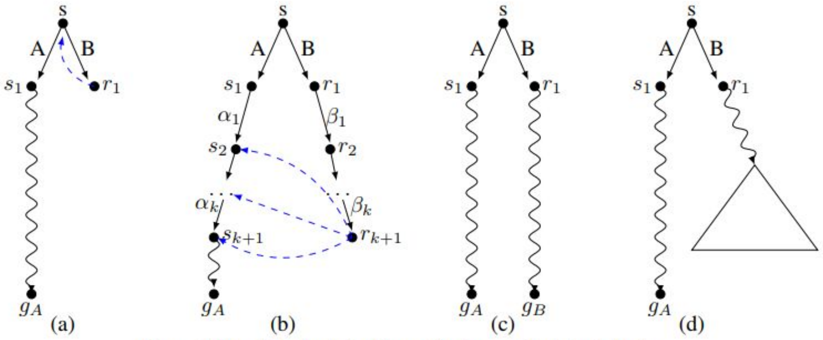
Rita Borgo, Michael Cashmore, Daniele Magazzeni: Towards Providing Explanations for AI Planner Decisions. CoRR abs/1810.06338 (2018)

Overview of Explanation in Different AI Fields (6)

- Planning and Scheduling

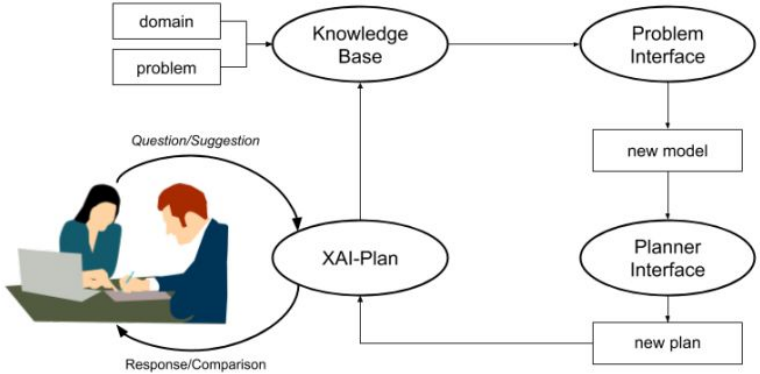
Explanation Type	R1	R2	R3	R4
Plan Patch Explanation / VAL	✗	✓	✗	✓
Model Patch Explanation	✓	✗	✓	✓
Minimally Complete Explanation	✓	✓	✗	?
Minimally Monotonic Explanation	✓	✓	✓	?
(Approximate) Minimally Complete Explanation	✗	✓	✗	✓

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



Human-in-the-loop Planning

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



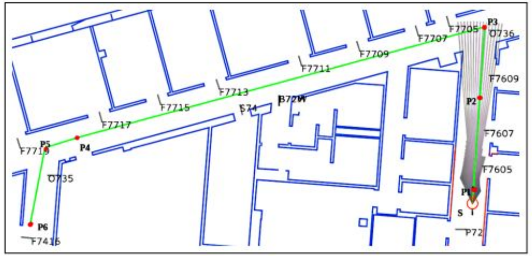
XAI Plan

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

(Manual) Plan Comparison

Overview of Explanation in Different AI Fields (7)

- Robotics



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

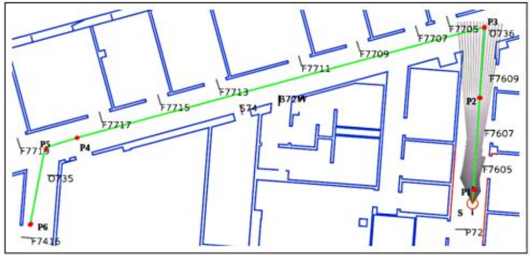
Narration of Autonomous Robot Experience

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

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

Overview of Explanation in Different AI Fields (7)

- Robotics



		Abstraction, A			
		Level 1	Level 2	Level 3	Level 4
Specificity, S	General Picture	Start and finish point of the complete route	Total distance and time taken for the complete route	Total distance and time taken for the complete route	Starting and ending land-mark 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 build-ing	Total distance and angles for subroute on each floor of each building	Starting and ending land-mark for subroute on each floor of each build-ing
	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 dis-tance and time taken for each edge of route	Turn direction at each point plus total distance for each edge of route	All landmarks encoun-tered 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?

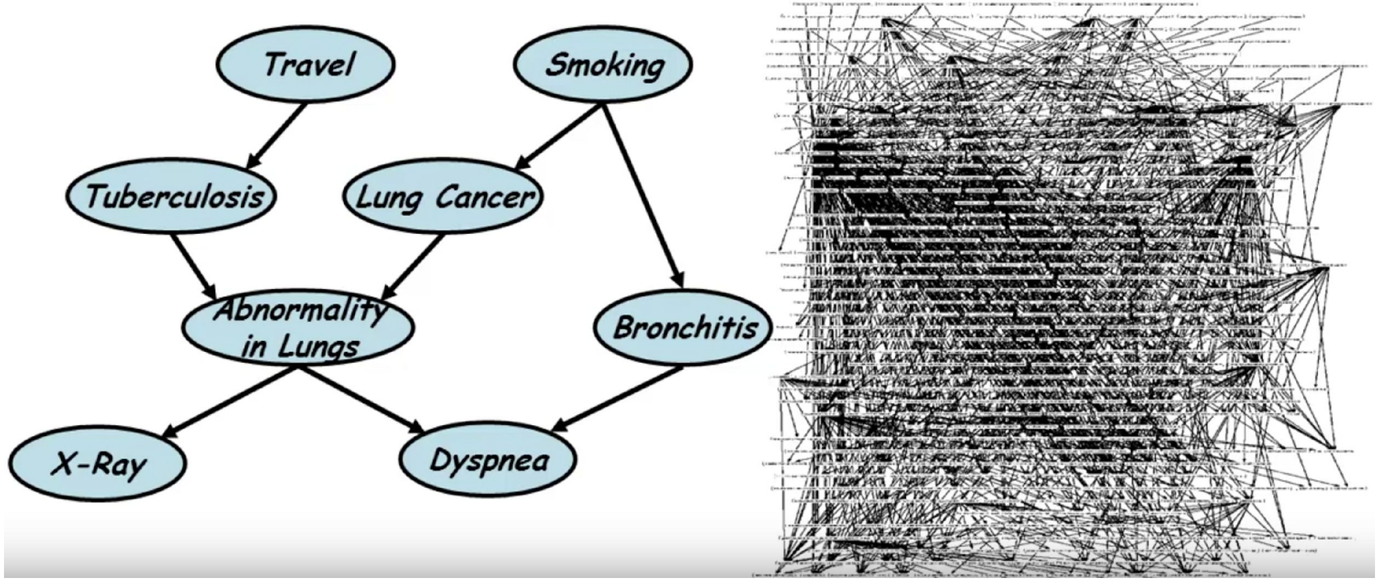
Robot: Here is the histogram of all training examples that reached this leaf. 80% of examples where this area was above 20 cm predicted the appropriate action to be "drive forward".

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)

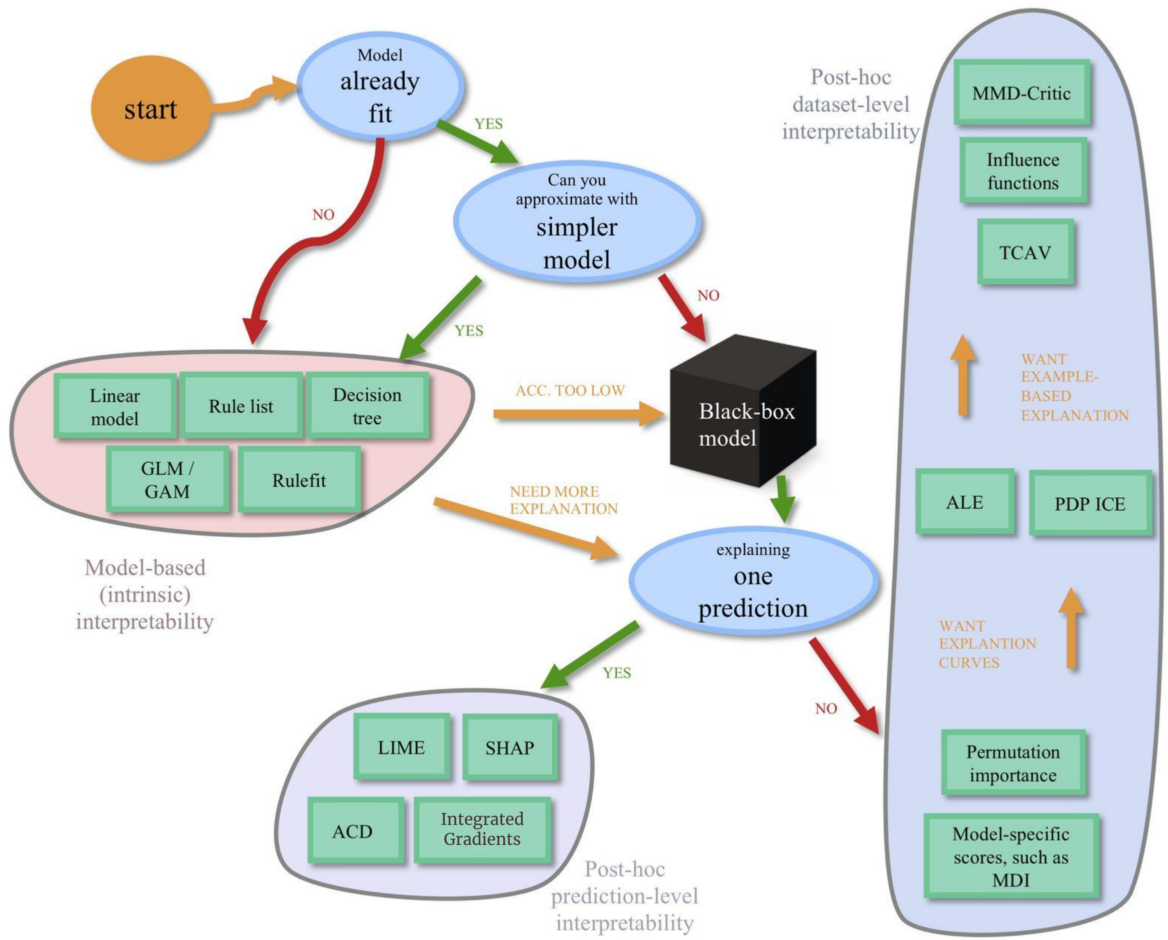
Achieving Explainable AI

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

Slide credit: https://twitter.com/chandan_singh96/status/1138811752769101825

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)

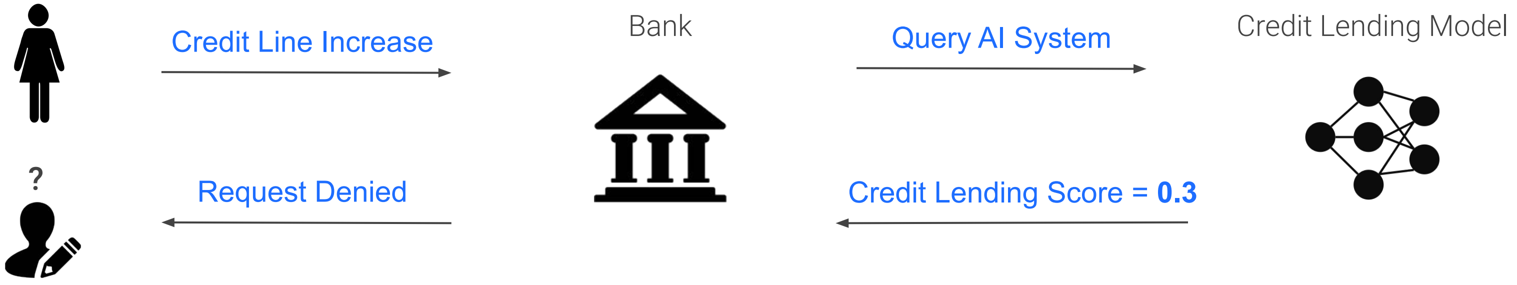
Top label: **"clog"**

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

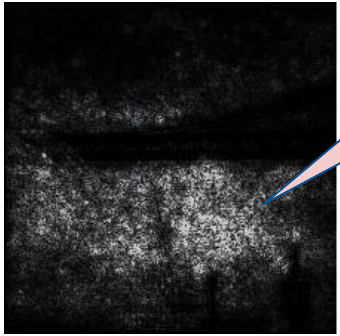
Cons:

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

Feature*Gradient

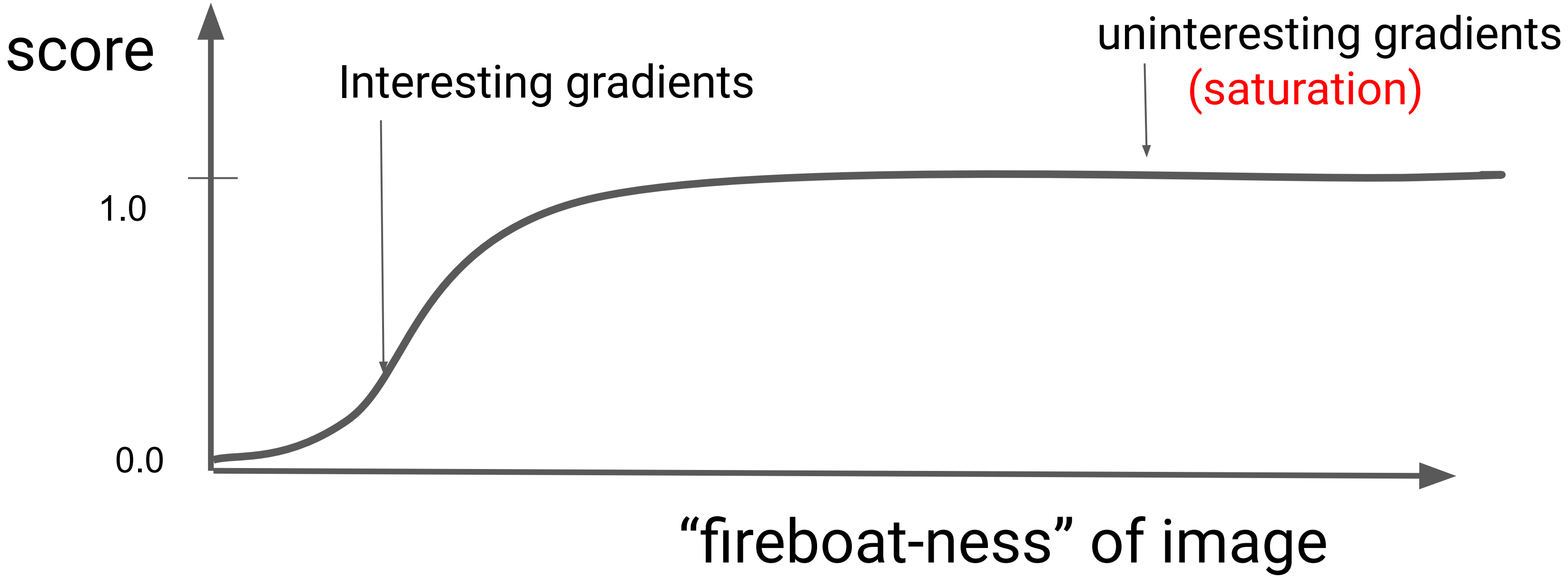
Attribution to a feature is feature value times gradient, i.e., $x_i * \partial y/\partial x_i$

- 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?

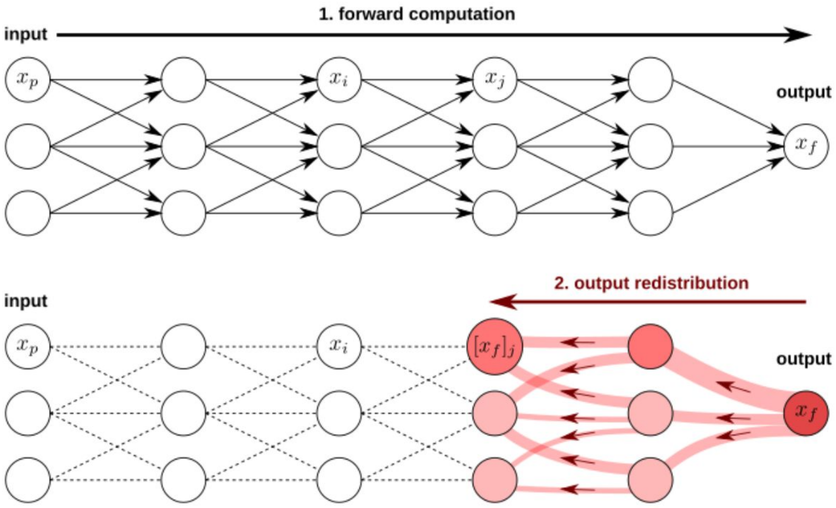
Local linear approximations can be too local



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}

Image credit heatmapping.org

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]

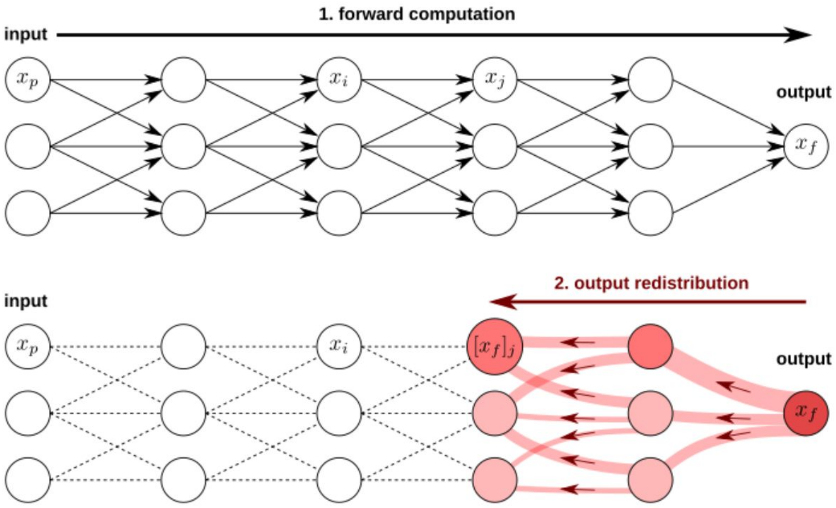


Image credit heatmapping.org

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]

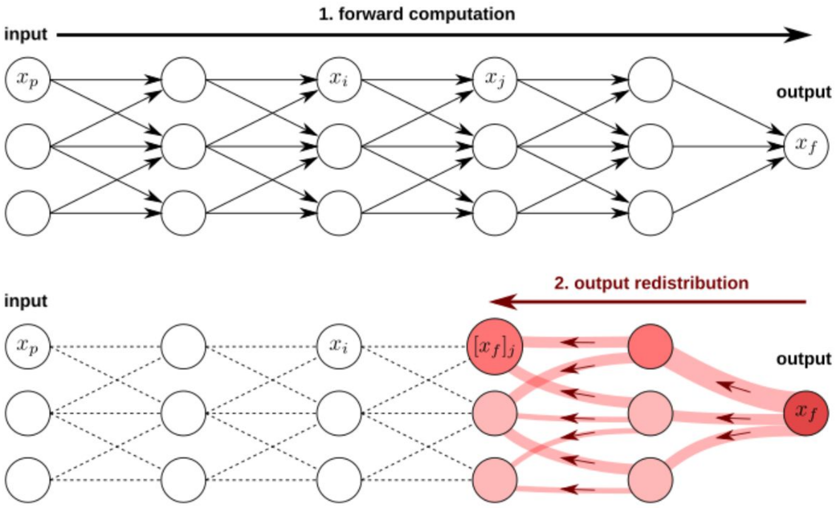


Image credit heatmapping.org

Pros:

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

Cons:

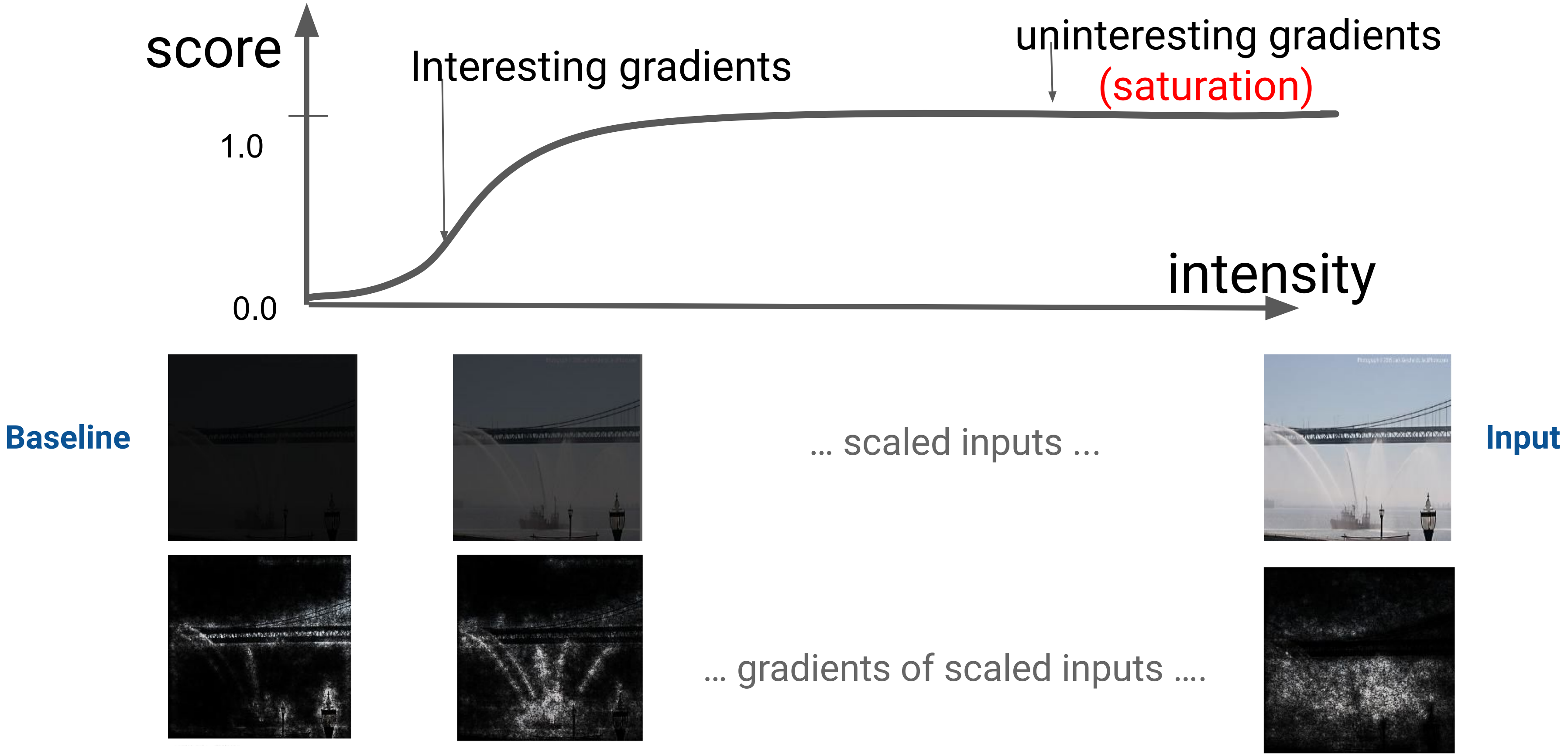
- Hard to implement, requires instrumenting the model
- **Often breaks implementation invariance**

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 Pr(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(input) - F(baseline)** in terms of input features

Another approach: gradients at many points



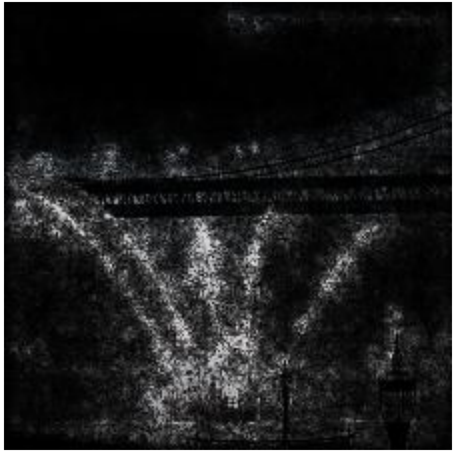
Integrated Gradients [ICML 2017]

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

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

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”?

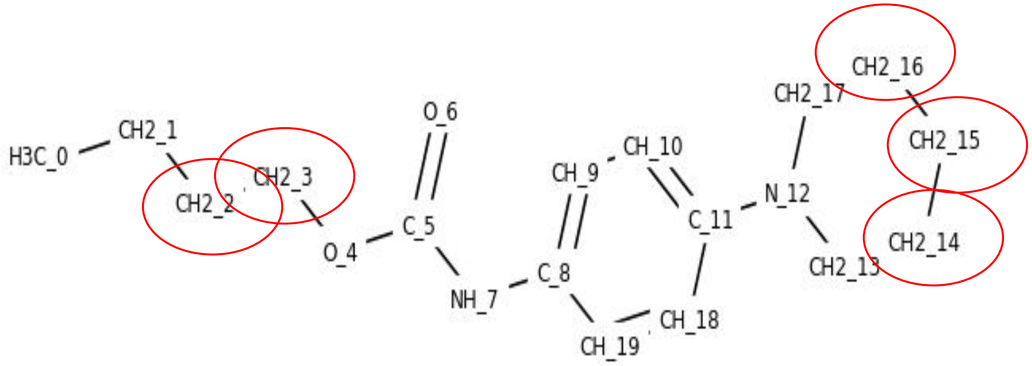
Original image

Integrated Gradients
(for label “clog”)

“Clog”

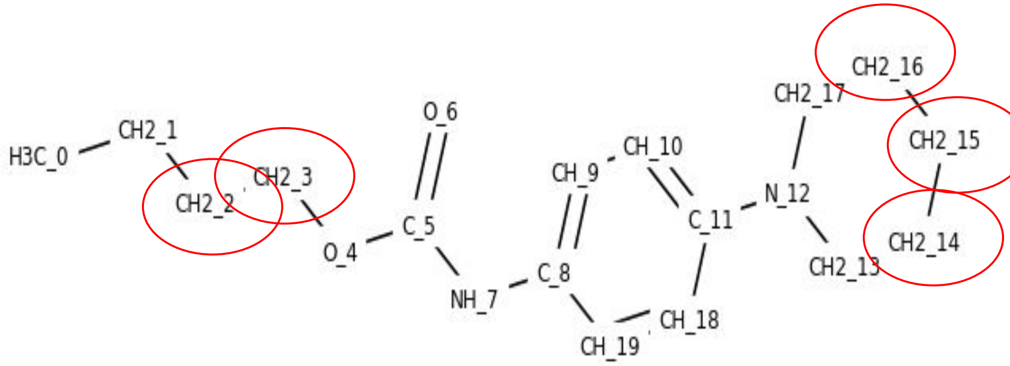
Detecting an architecture bug

- 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

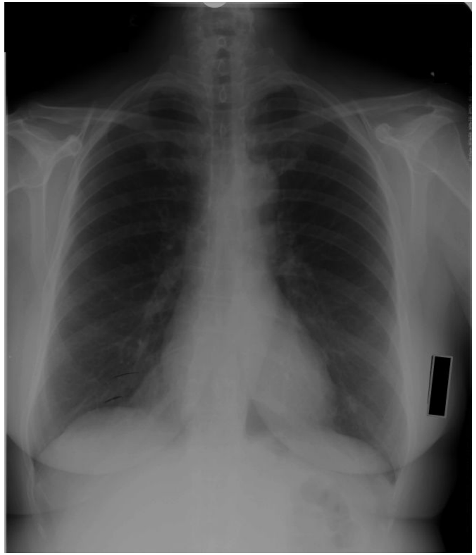


- **Bug**: The architecture had a bug due to which the convolved bond features did not affect the prediction!

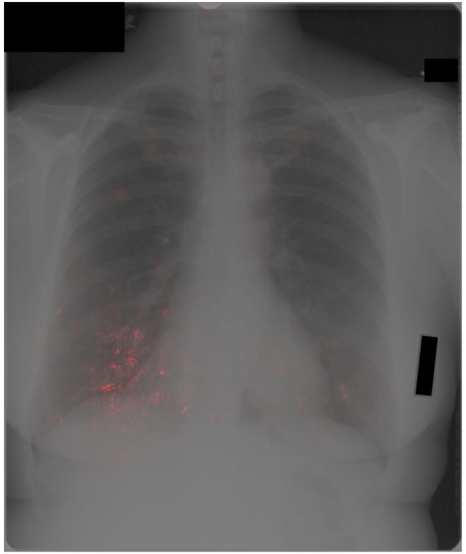
Detecting a data issue

- Deep network predicts various diseases from chest x-rays

Original image



Integrated gradients
(for top label)

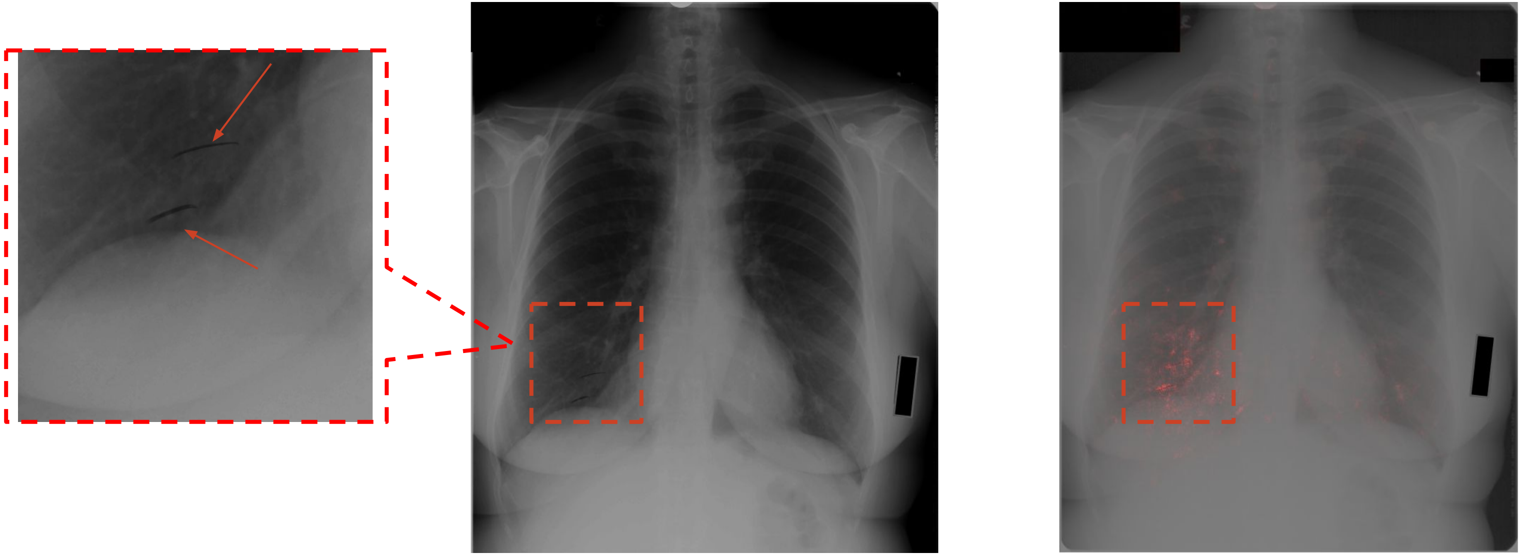


Detecting a data issue

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

Original image

Integrated gradients (for top label)



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 U {i}) - v(S))
  - Shapley value for a player is a **specific weighted aggregation of its marginal** over all possible subsets of other players

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

(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, <absent>, x_3, …, <absent>)**?

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

Preprint: The Explanation Game: Explaining Machine Learning Models with Cooperative Game Theory

Computing Shapley Values

Exact Shapley value computation is exponential in the number of features

- Shapley values can be expressed as an expectation of marginals

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

  - 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:

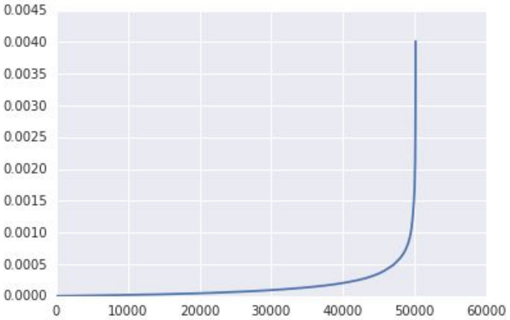
- A model that predicts TRUE when there are **even number** of black pixels and FALSE otherwise

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)

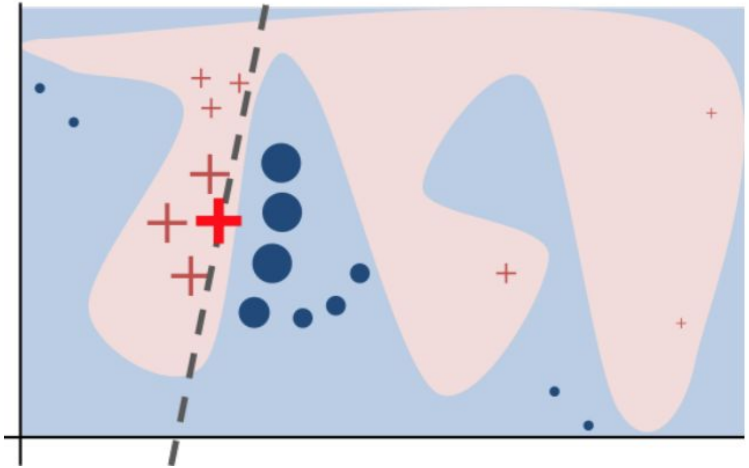
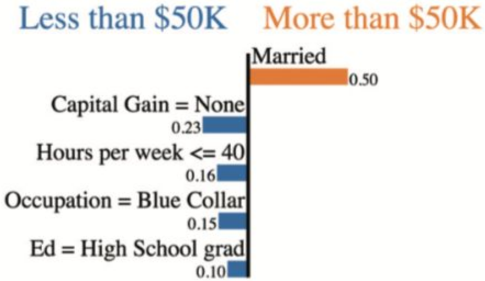


Figure credit: Ribeiro et al. KDD 2016

$28 < \text{Age} \leq 37$
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 ≤ 40.00
Country = United-States
$P(\textbf{Salary} > \$50\textbf{K}) = 0.57$

(a) Instance and prediction



(b) LIME explanation

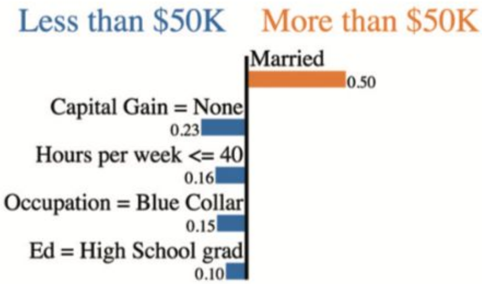
Figure credit: Anchors: High-Precision Model-Agnostic Explanations. Ribeiro et al. AAAI 2018

Anchors

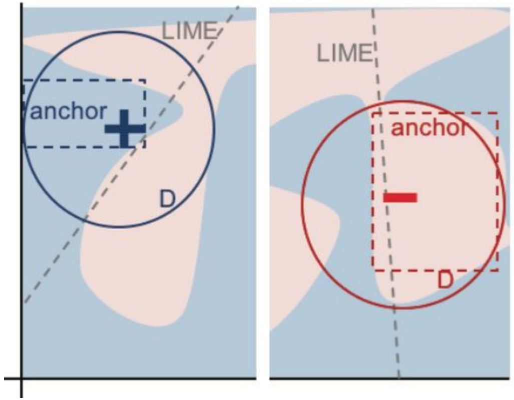
$28 < \text{Age} \leq 37$
Workclass = Private
Education = High School grad
Marital Status = Married
Occupation = Blue-Collar
Relationship = Husband
Race = White
Sex = Male
Capital Gain = None
Capital Loss = Low
Hours per week $\leq$ 40.00
Country = United-States

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

(a) Instance and prediction



(b) LIME explanation



IF Country = United-States **AND** Capital Loss = Low
AND Race = White **AND** Relationship = Husband
AND Married **AND** $28 < \text{Age} \leq 37$
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

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

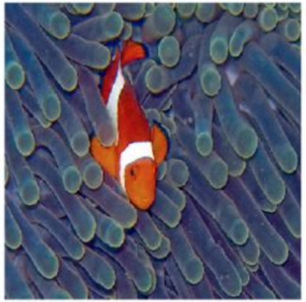
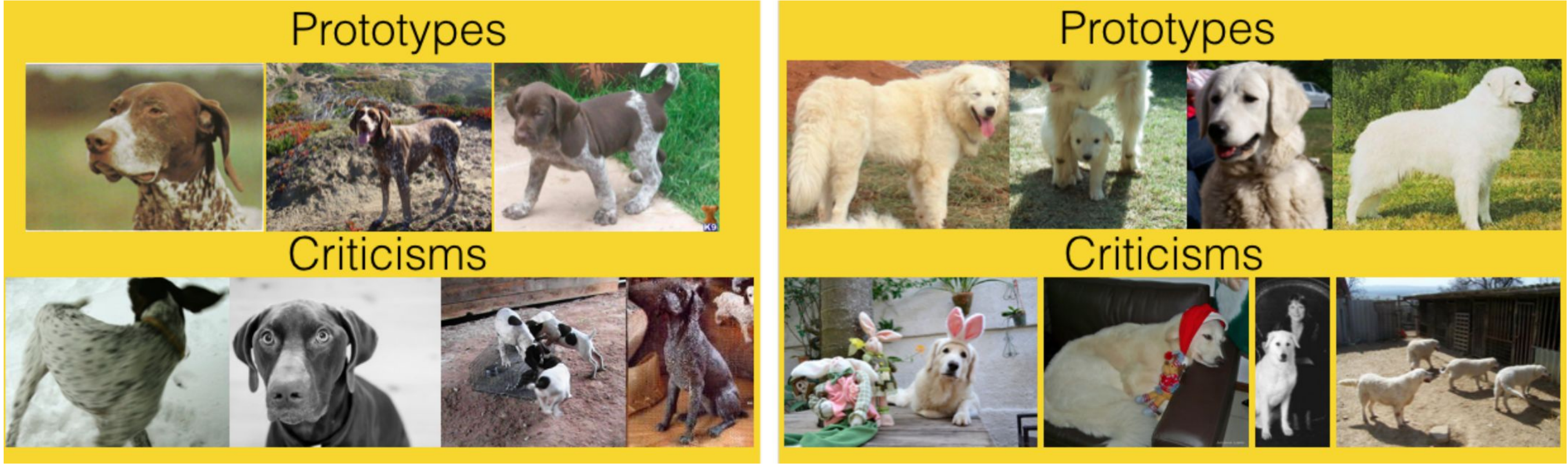


Figure credit: Understanding Black-box Predictions via Influence Functions. Koh and Liang. ICML 2017

Example based Explanations



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

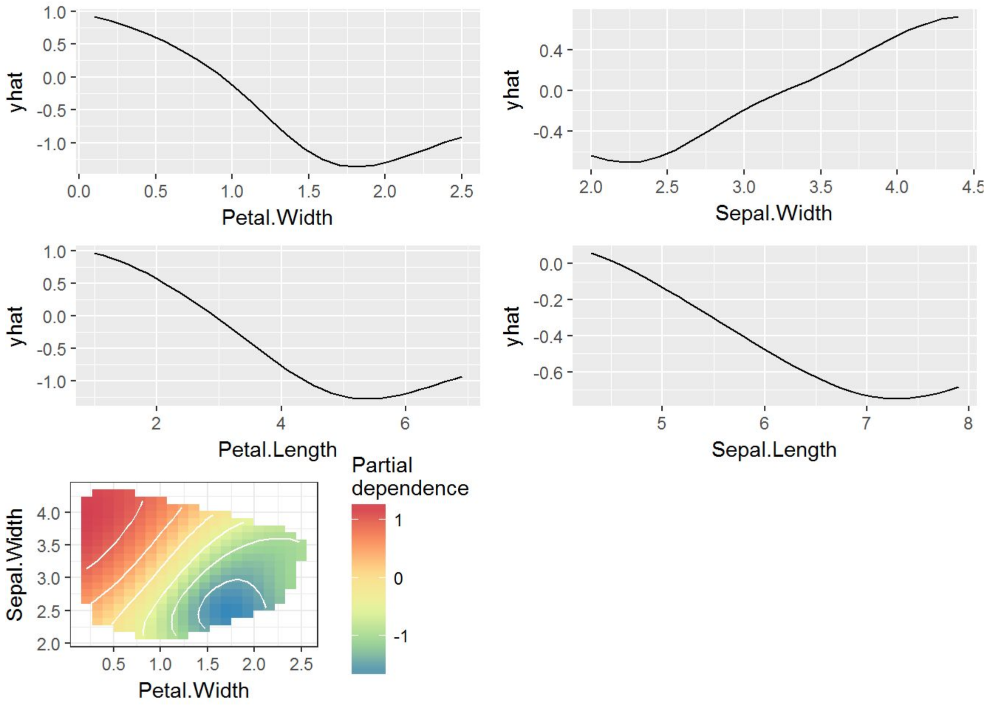
- Prototypes: Representative of all the training data.
- Criticisms: Data instance that is not well represented by the set of prototypes.

Figure credit: Examples are not Enough, Learn to Criticize! Criticism for Interpretability. Kim, Khanna and Koyejo. NIPS 2016

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.

	RDSpend	Administration	MarketingSpend	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**

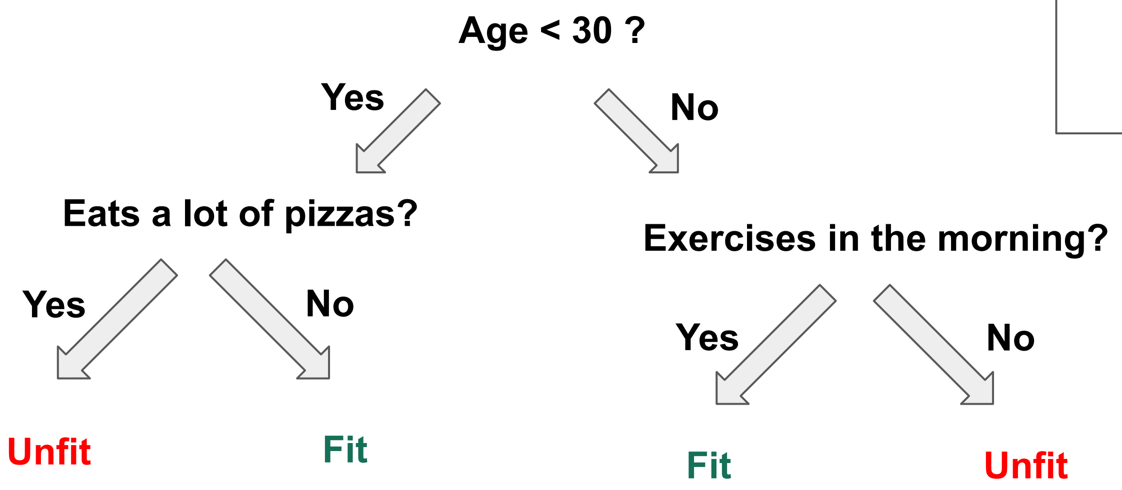
- **Individual prediction explanations** in terms of input features, influential examples, concepts, local decision rules
- **Global prediction explanations** in terms of entire model in terms of partial dependence plots, global feature importance, global decision rules

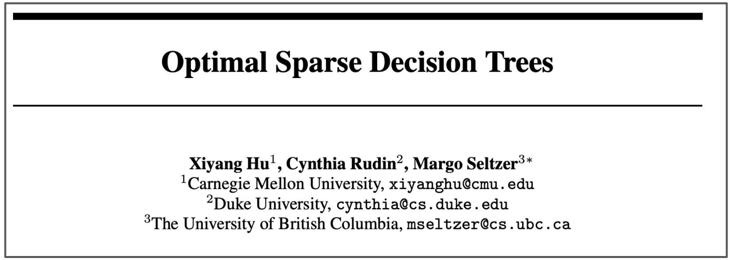
Approach 2: **Build an interpretable model**

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

Decision Trees

Is the person fit?





Optimal Sparse Decision Trees

Xiyang Hu[1], **Cynthia Rudin**[2], **Margo Seltzer**[3*]
[1]Carnegie Mellon University, xiyanghu@cmu.edu
[2]Duke University, cynthia@cs.duke.edu
[3]The 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 $\geq$ 0.1, **then** Asthma

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

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

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

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

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

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

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

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

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

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

Figure credit: Interpretable Decision Sets: A Joint Framework for Description and Prediction, Lakkaraju, Bach, Leskovec

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 $\cdots$, *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 $\geq$ 50, **then** Lung Cancer

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

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

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

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

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

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

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

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

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

Else Diabetes

Figure credit: Interpretable Decision Sets: A Joint Framework for Description and Prediction, Lakkaraju, Bach, Leskovec

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 ≥ 60	THEN malignancy risk is	85.22%	230
ELSE IF	SpiculatedMargin AND Age ≥ 45	THEN malignancy risk is	78.13%	64
ELSE IF	IllDefinedMargin AND Age ≥ 60	THEN malignancy risk is	69.23%	39
ELSE IF	IrregularShape	THEN malignancy risk is	63.40%	153
ELSE IF	LobularShape AND Density ≥ 2	THEN malignancy risk is	39.68%	63
ELSE IF	RoundShape AND Age ≥ 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

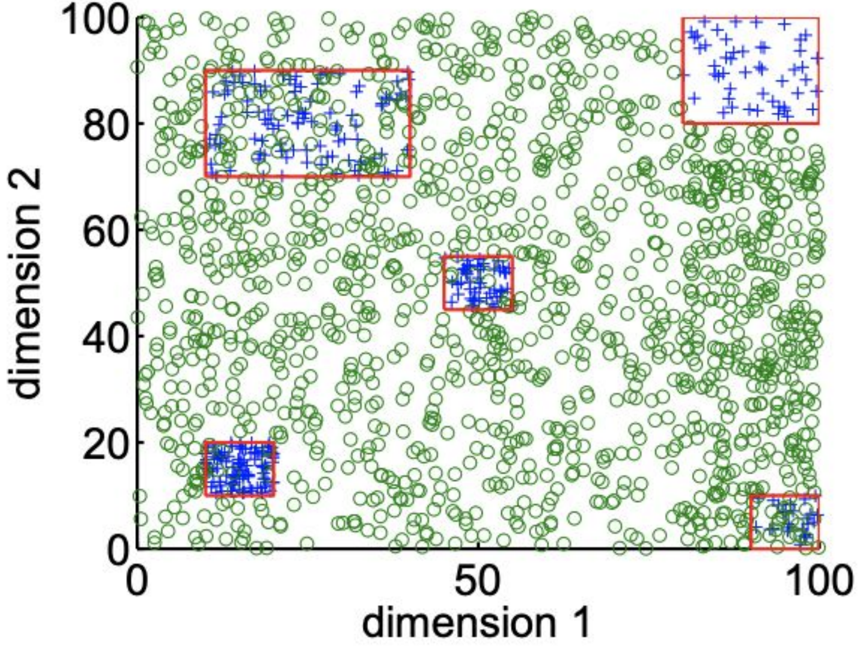


Figure credit: Box Drawings for Learning with Imbalanced. Data Siong Thye Goh and Cynthia Rudin

Supersparse Linear Integer Models for Optimized Medical Scoring Systems

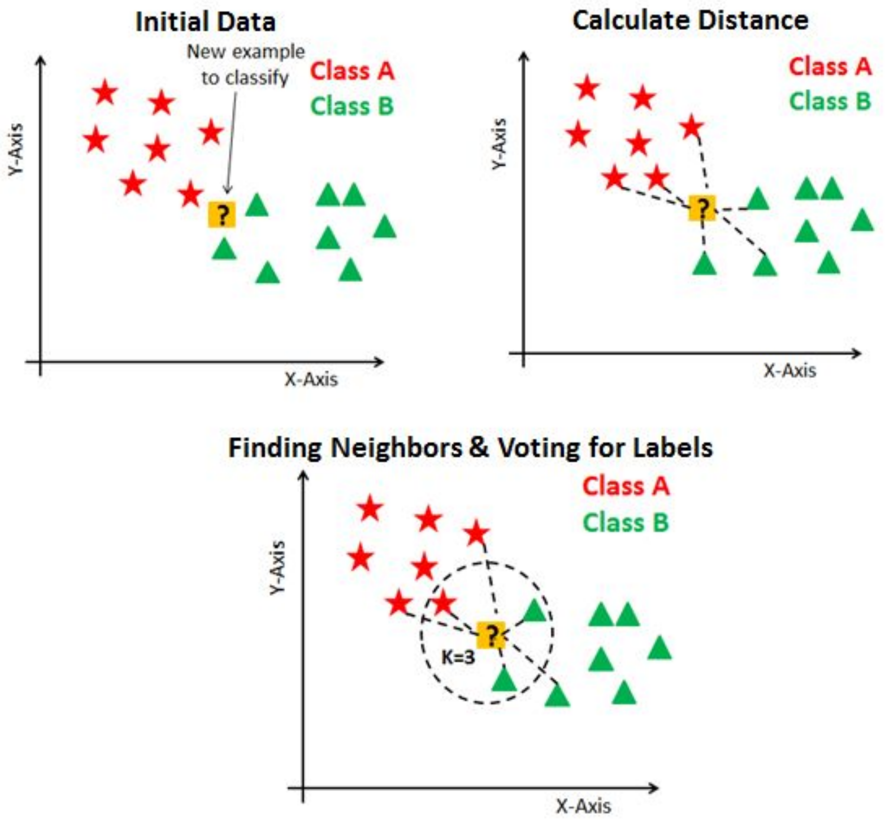
PREDICT PATIENT HAS OBSTRUCTIVE SLEEP APNEA IF SCORE > 1

1.	*age* ≥ 60	4 points	$\cdots\cdots$
2.	*hypertension*	4 points	$+ \quad \cdots\cdots$
3.	*body mass index* ≥ 30	2 points	$+ \quad \cdots\cdots$
4.	*body mass index* ≥ 40	2 points	$+ \quad \cdots\cdots$
5.	*female*	-6 points	$+ \quad \cdots\cdots$
	ADD POINTS FROM ROWS 1 – 5	**SCORE**	$= \quad \cdots\cdots$

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

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

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

Explainable Machine Learning

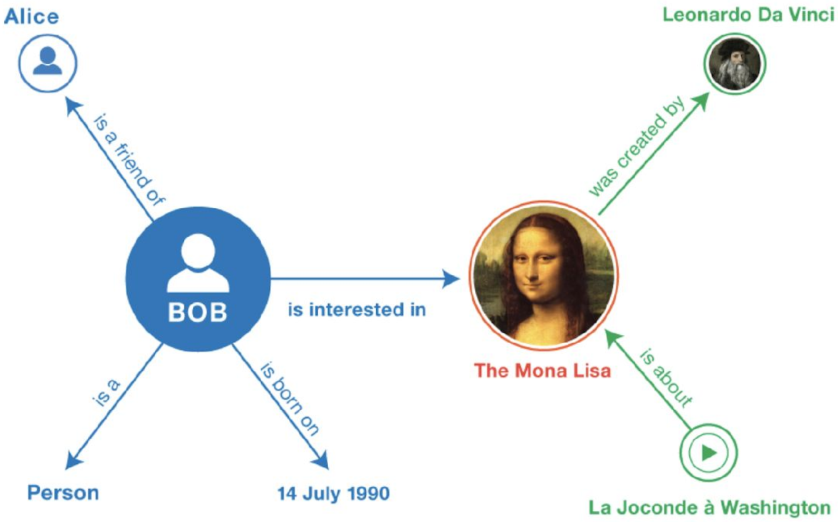
(from a Knowledge Graph Perspective)

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

Knowledge Graph (1)

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

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



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

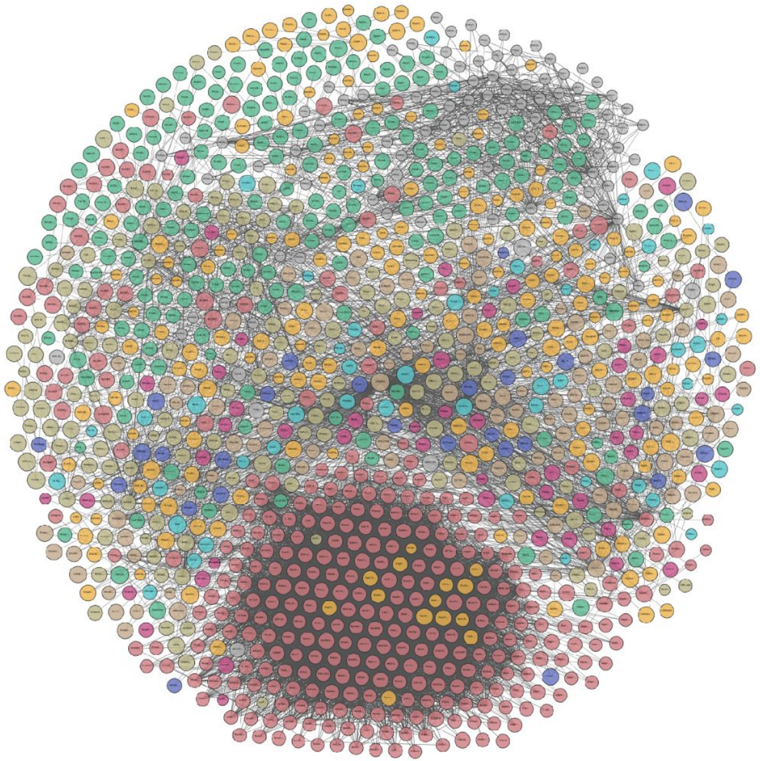
Knowledge Graph (2)

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

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

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

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



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

Knowledge Graph Construction

Knowledge Graph construction methods can be classified in:

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

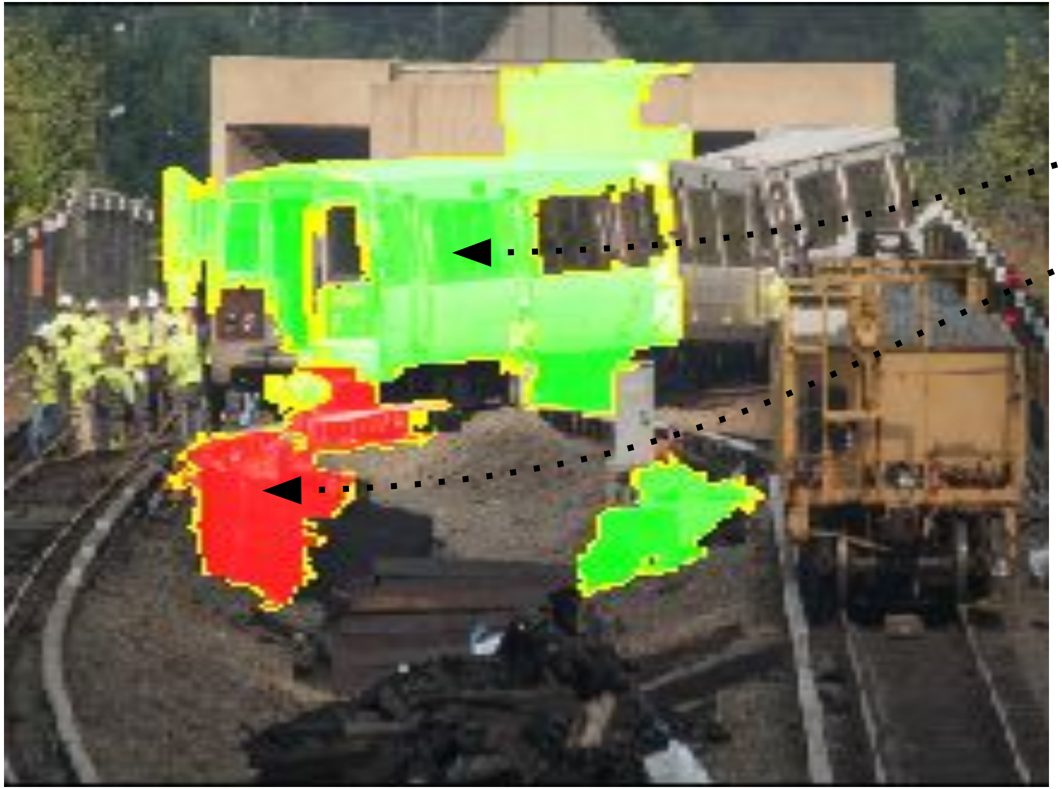
Coverage is an issue:

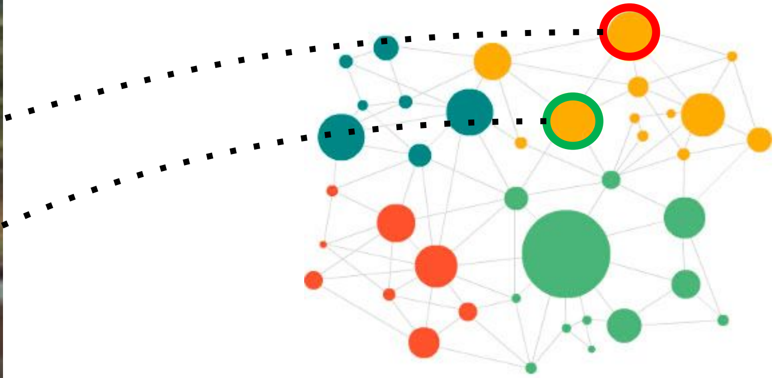
- **Freebase** (40M entities) - 71% of persons without a birthplace, 75% without a nationality, even worse for other relation types [Dong et al. 2014]
- **DBpedia** (20M entities) - 61% of persons without a birthplace, 58% of scientists missing why they are popular [Krompaß et al. 2015]

Relational Learning can help us overcoming these issues.

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

Knowledge Graph in Machine Learning (1)



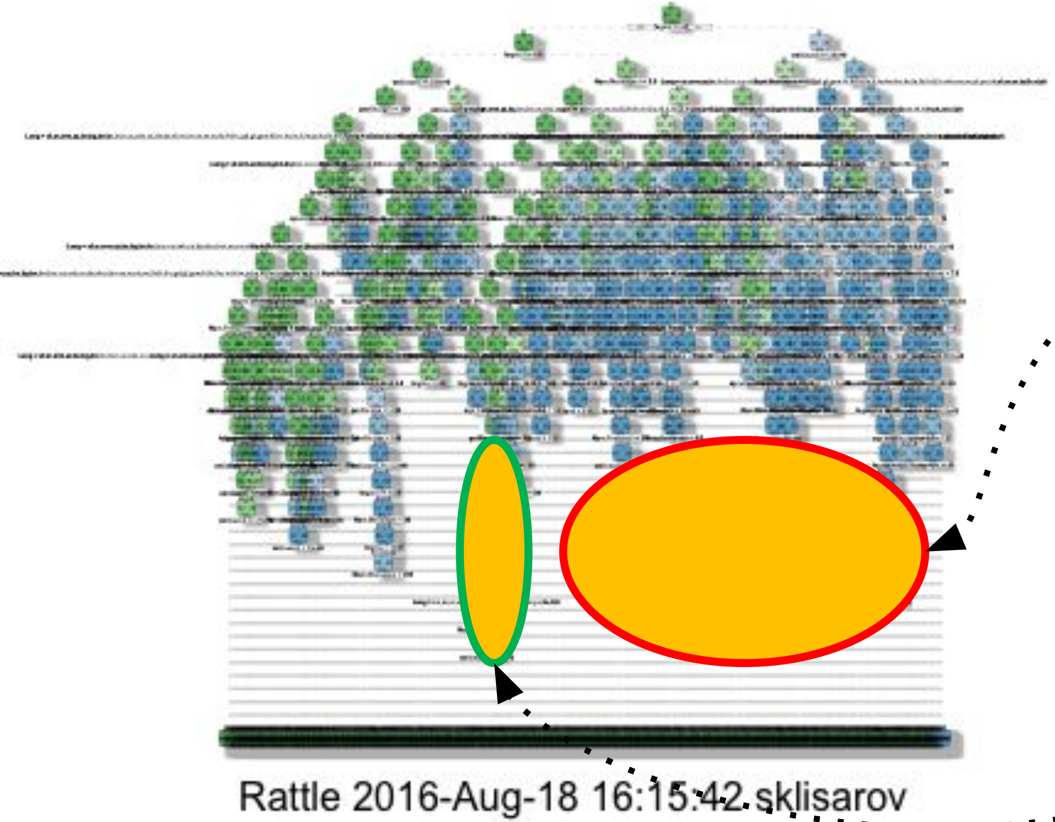


Augmenting (input) features with more semantics such as knowledge graph embeddings / entities

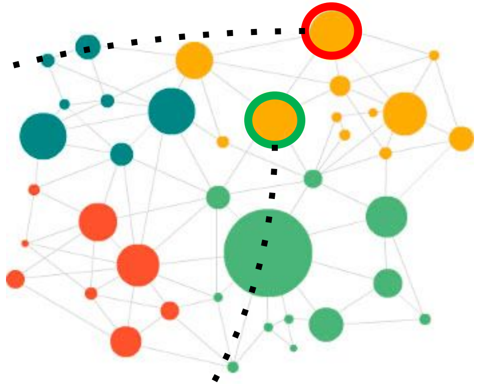
https://stats.stackexchange.com/questions/230581/decision-tree-too-large-to-interpret

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

Knowledge Graph in Machine Learning (2)



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



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

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

Knowledge Graph in Machine Learning (3)

Input Layer

Hidden Layer

Output Layer

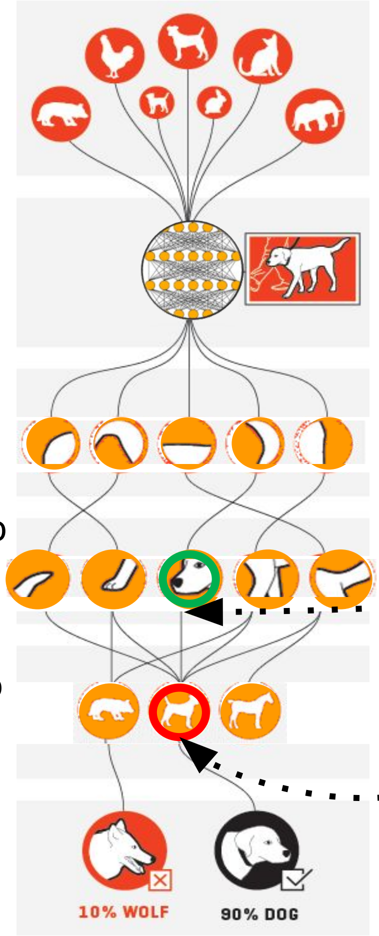
Training Data

Input (unlabeled image)

Neurons respond to simple shapes

Neurons respond to more complex structures

Neurons respond to highly complex, abstract concepts

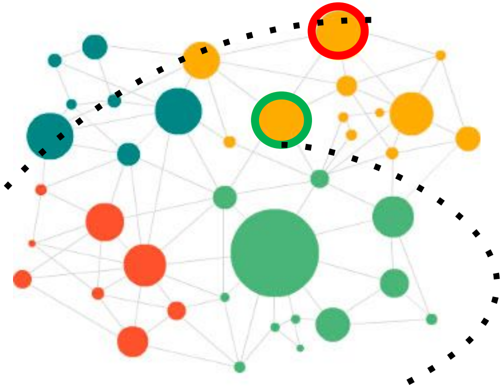


1st Layer

2nd Layer

nth Layer

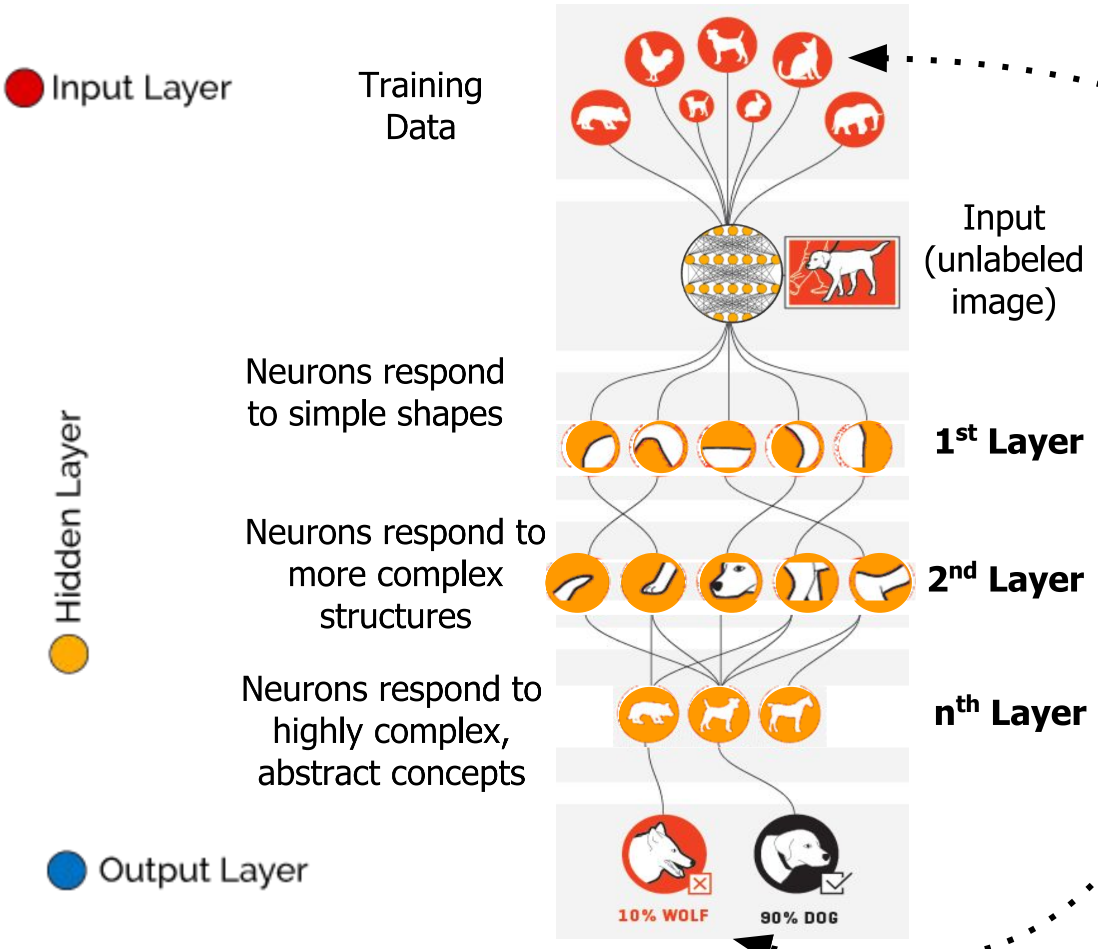
Low-level features to high-level features

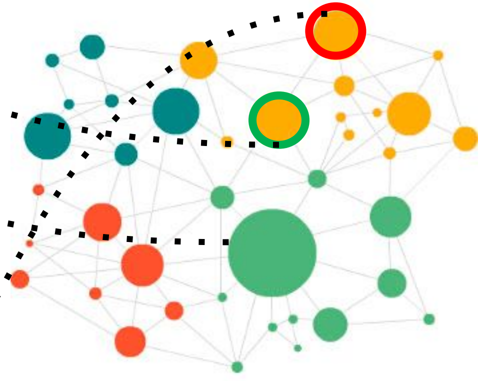


Augmenting (intermediate) features with more semantics such as knowledge graph embeddings / entities

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

Knowledge Graph in Machine Learning (4)

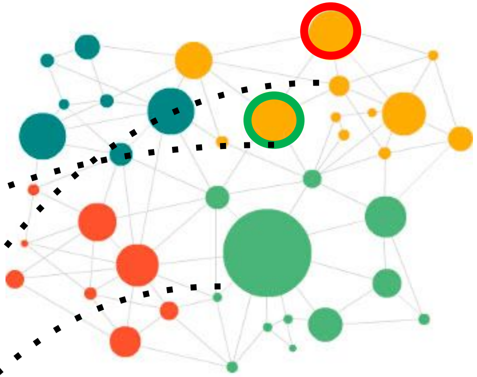




Augmenting (input, intermediate) features – output relationship with more semantics to capture causal relationship

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

Knowledge Graph in Machine Learning (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

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

Knowledge Graph in Machine Learning (6)

"How to explain transfer learning with appropriate knowledge representation?

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

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

Knowledge-Based Transfer Learning Explanation

Jiaoyan Chen
Department of Computer Science
University of Oxford, UK

Freddy Lecue
INRIA, France
Accenture Labs, Ireland

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

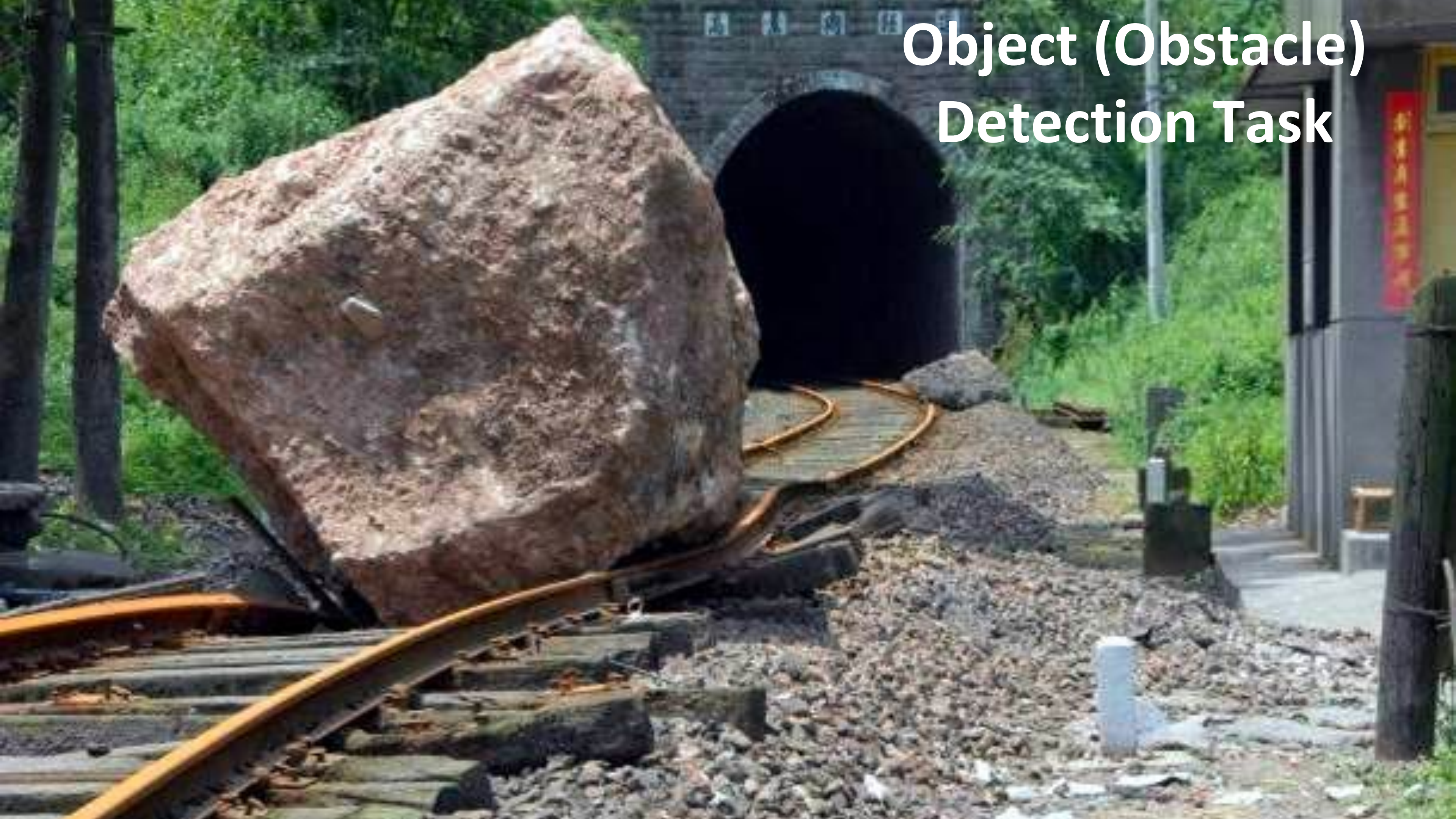
Ian Horrocks
Department of Computer Science
University of Oxford, UK

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

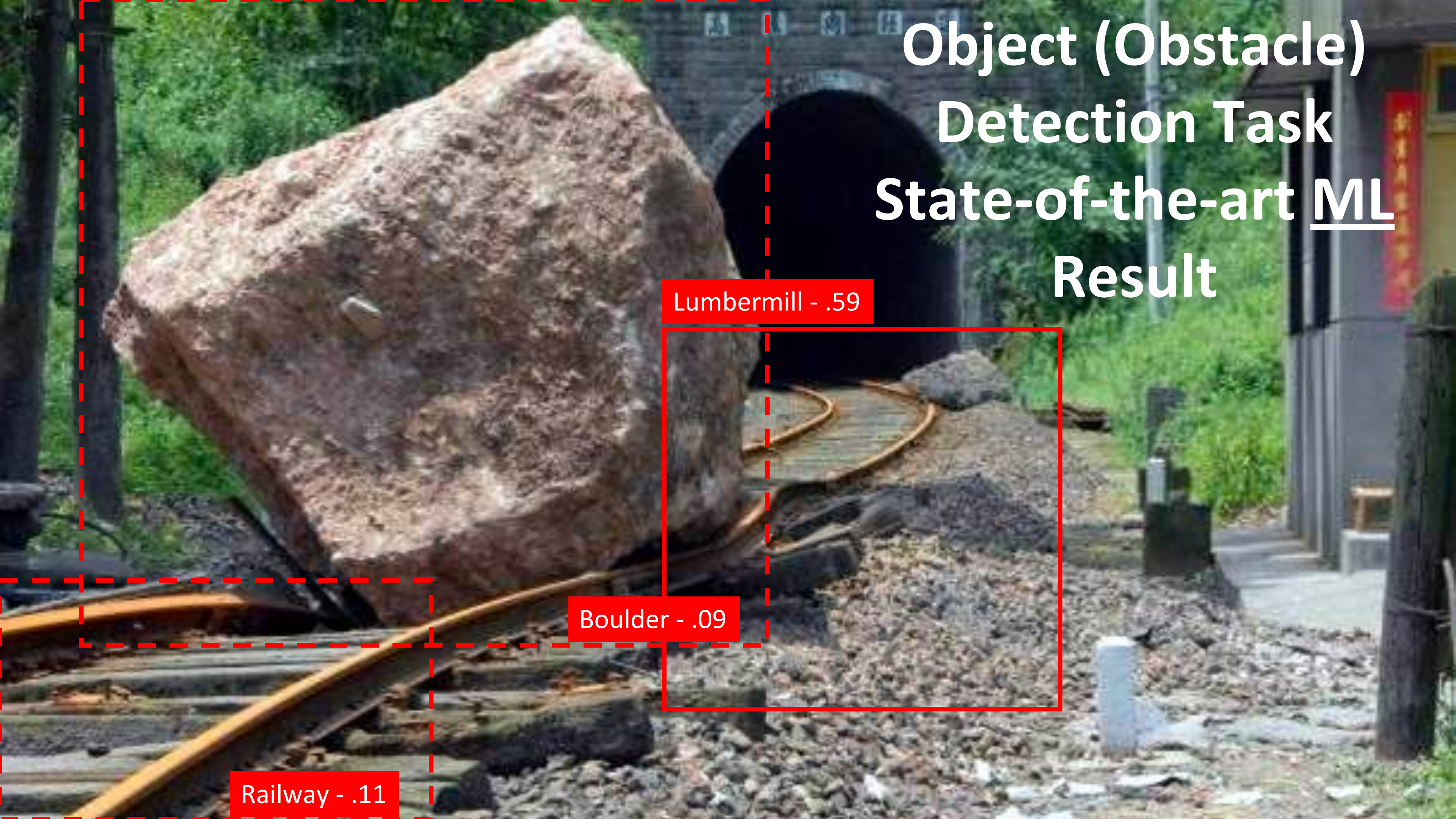
How Does
it
Work
in Practice?

State of the Art Machine Learning Applied to Critical Systems

Object (Obstacle) Detection Task




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


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

State of the Art XAI Applied to Critical Systems


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

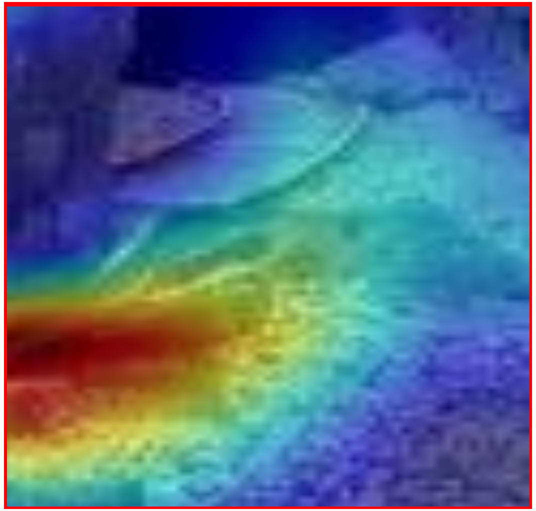
Unfortunately, this is of NO use for a human behind the system

Let’s stay back

Why this Explanation?
(meta explanation)

After Human Reasoning…

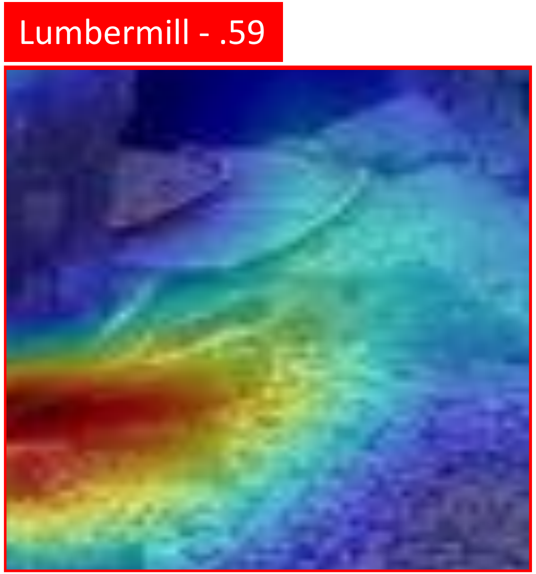
Lumbermill - .59



DBpedia   Browse using   Formats   Faceted Browser   Sparql Endpoint

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

What is missing?



DBpedia Browse using Formats Faceted Browser Sparql Endpoint

About: Boulder

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

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

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

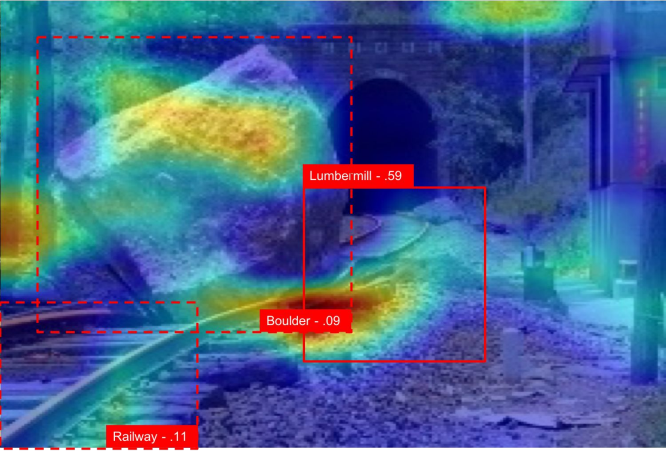
DBpedia Browse using Formats Faceted Browser Sparql Endpoint

About: Rail transport

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

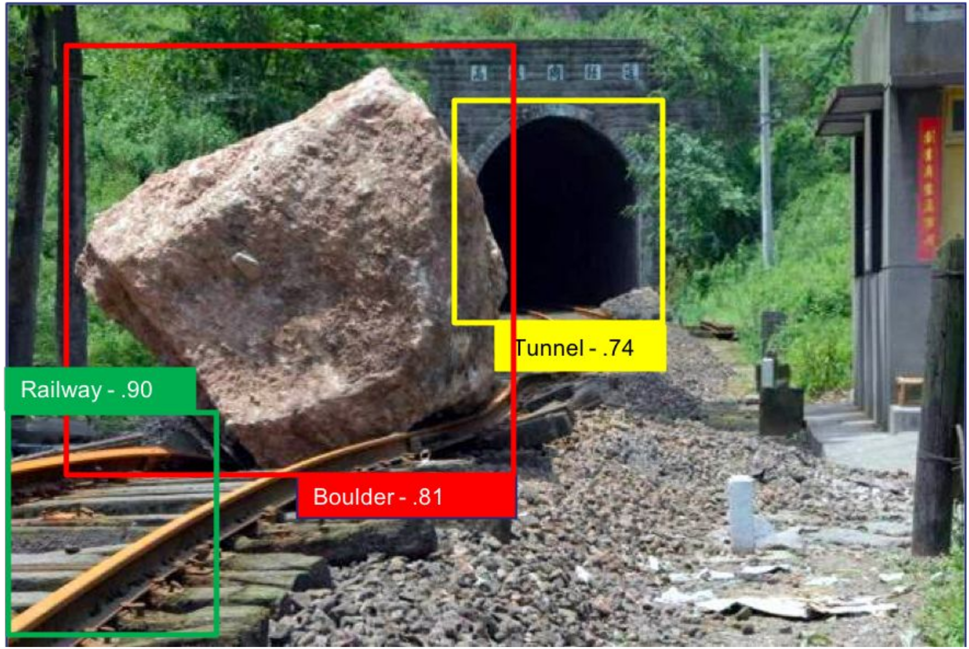
Rail transport is a means of conveyance of passengers and goods on wheeled vehicles running on rails, also known as tracks. It is also commonly referred to as train transport. In contrast to road transport, where vehicles run on a prepared flat surface, rail vehicles (rolling stock) are directionally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

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

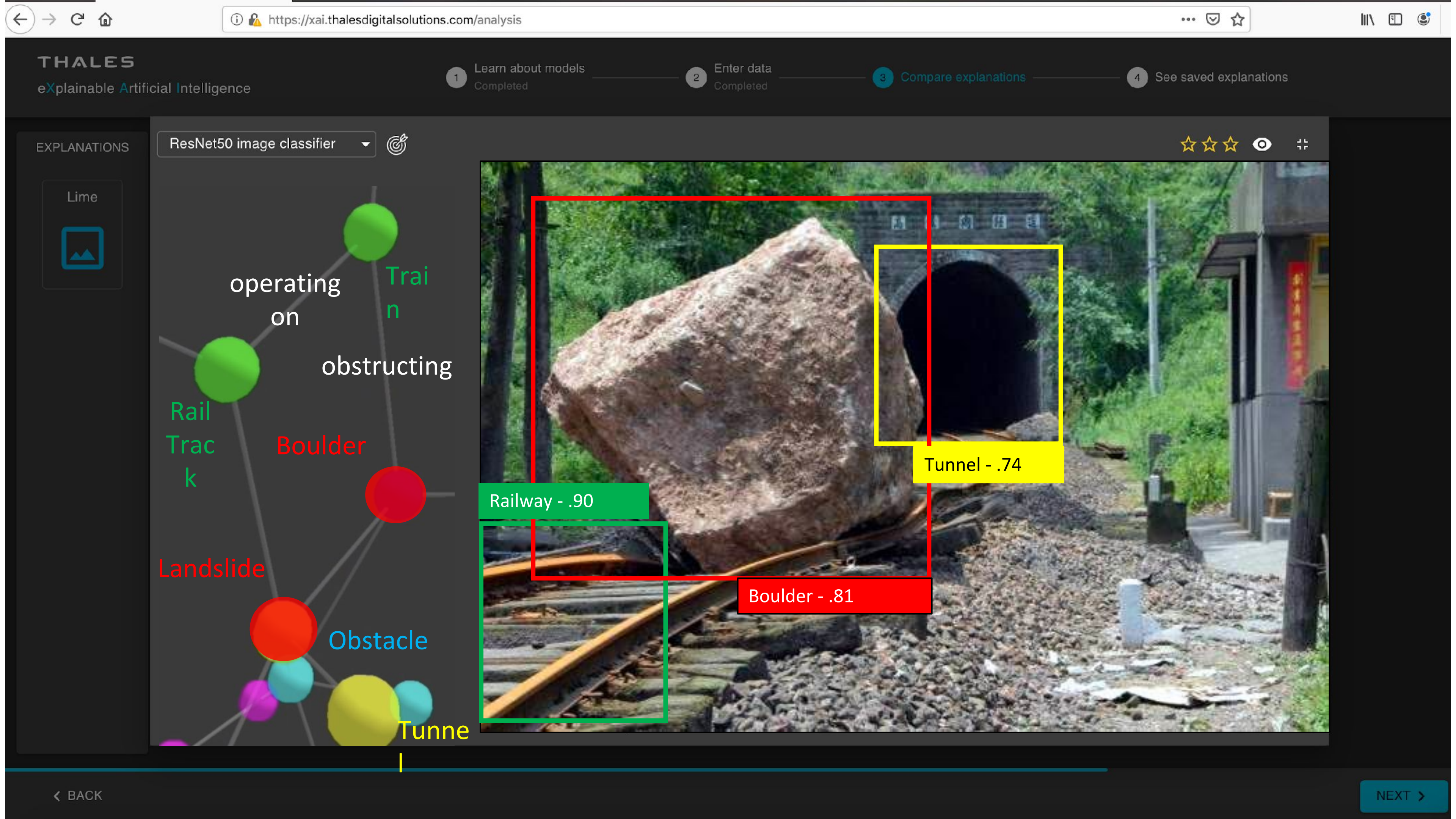


- **Hardware**: **High performance, scalable, generic** (to different FGPA family) **& portable** CNN dedicated **programmable** processor implemented on an FPGA for **real-time embedded inference**
- **Software**: Knowledge graph extension of object detection

Transitioning

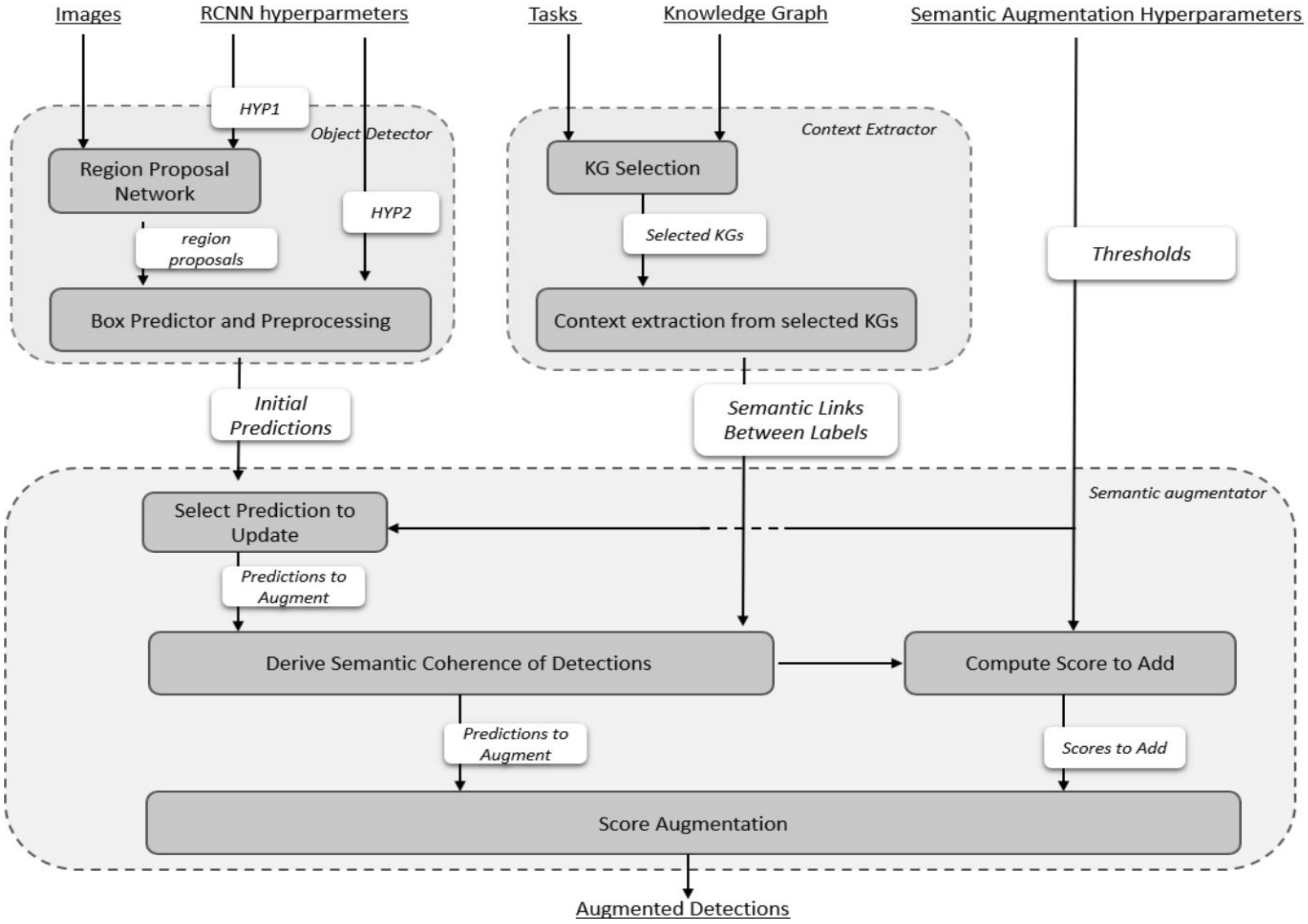


This is an **Obstacle: Boulder** obstructing the train: XG142-R on **Rail_Track** from City: Cannes to City: Marseille at **Location: Tunnel VIX** due to **Landslide**



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

Knowledge Graph in Machine Learning - An Implementation



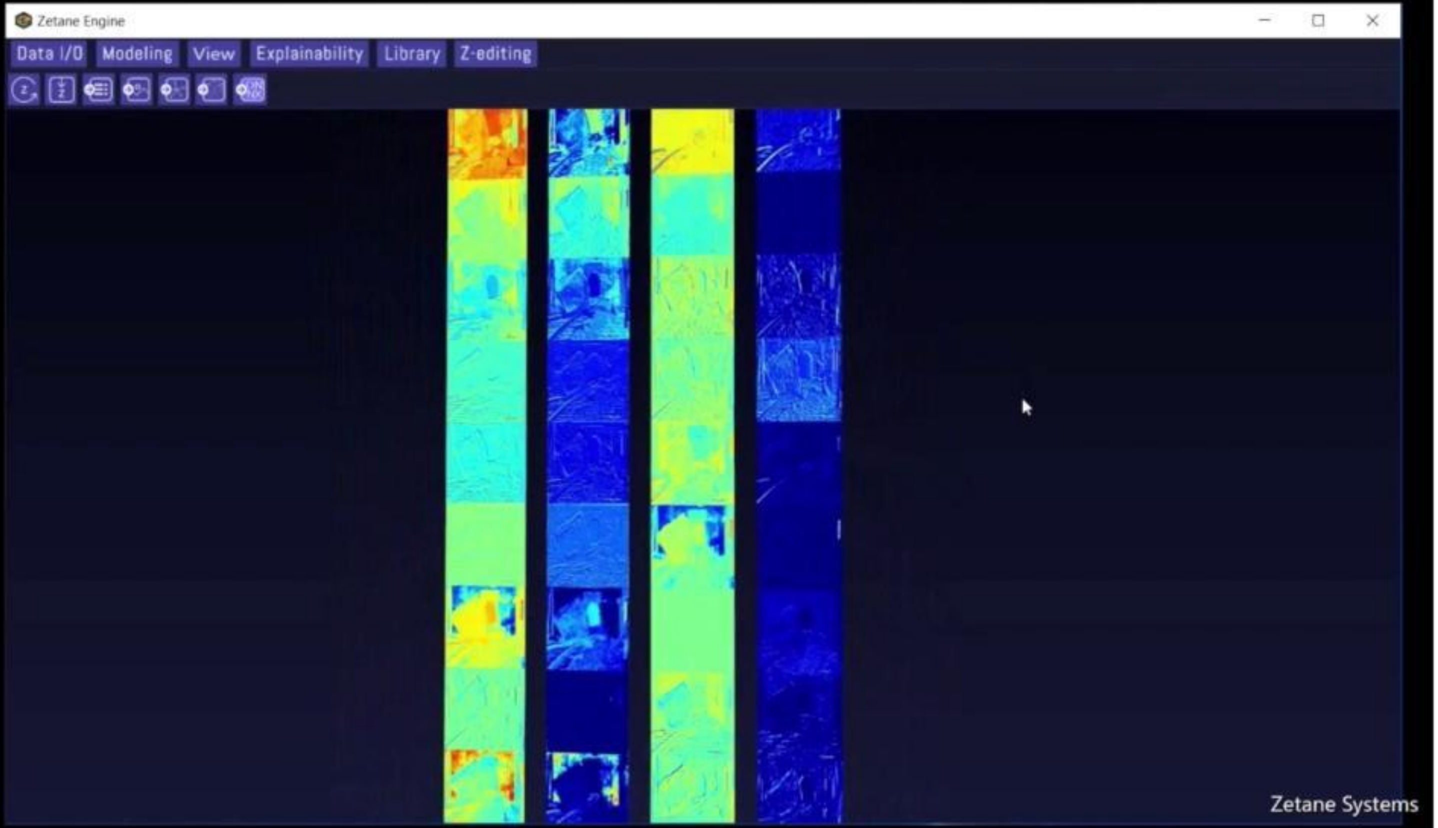
Freddy Lécué, Jiaoyan Chen, Jeff Z. Pan, Huajun Chen: Augmenting Transfer Learning with Semantic Reasoning. IJCAI 2019: 1779-1785

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

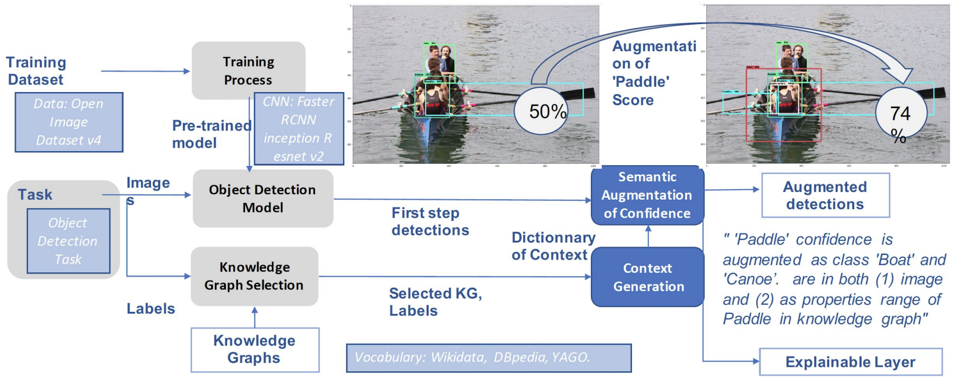
Jiaoyan Chen, Freddy Lécué, Jeff Z. Pan, Ian Horrocks, Huajun Chen: Knowledge-Based Transfer Learning Explanation. KR 2018: 349-358

XAI Tools on Applications, Lessons Learnt and Research Challenges



Zetane Engine
Data I/O
Modeling
View
Explainability
Library
Z-editing
Zetane Systems

Explainable Boosted Object Detection – Industry Agnostic



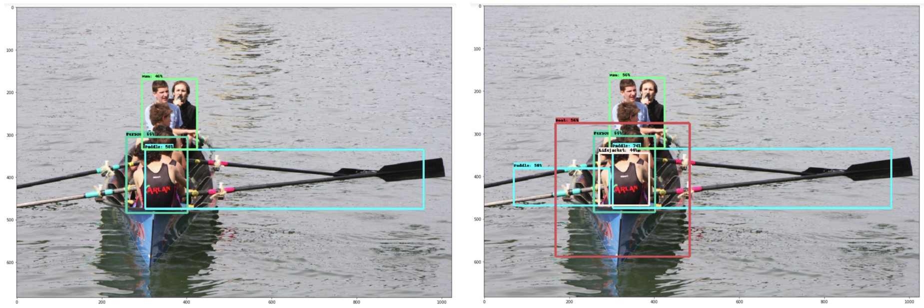


Fig. 2. Left image: results from baseline Faster RCNN: Paddle: 50% confidence, Person: 66%, Man: 46%. Right image: results from the semantic augmentation: **Paddle: 74%** confidence, Person: 66%, **Man: 56%**, **Boat: 58% with explanation: Person, Paddle, Water as part of the context in the image and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

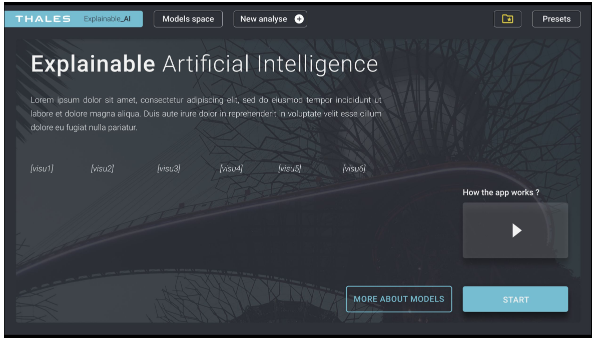
Challenge: Object detection is usually performed from a large portfolio of Artificial Neural Networks (ANNs) architectures trained on large amount of labelled data. Explaining object detections is rather difficult due to the high complexity of the most accurate ANNs.

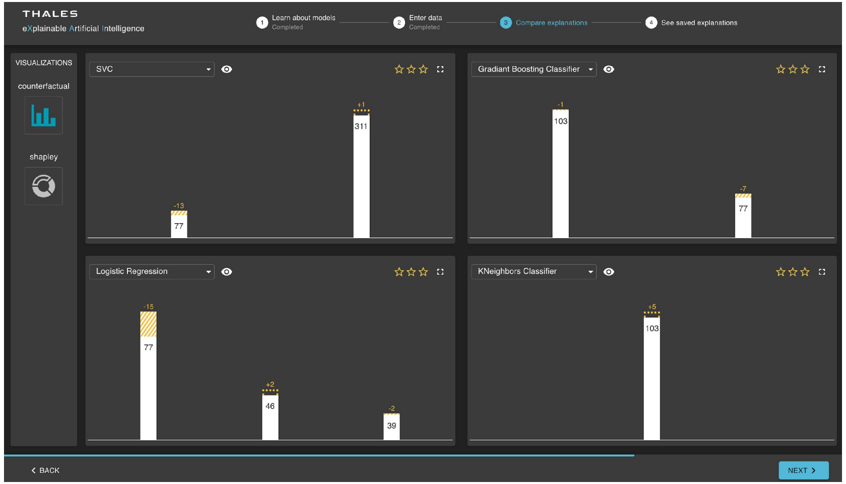
AI Technology: Integration of AI related technologies i.e., Machine Learning (Deep Learning / CNNs), and knowledge graphs / linked open data.

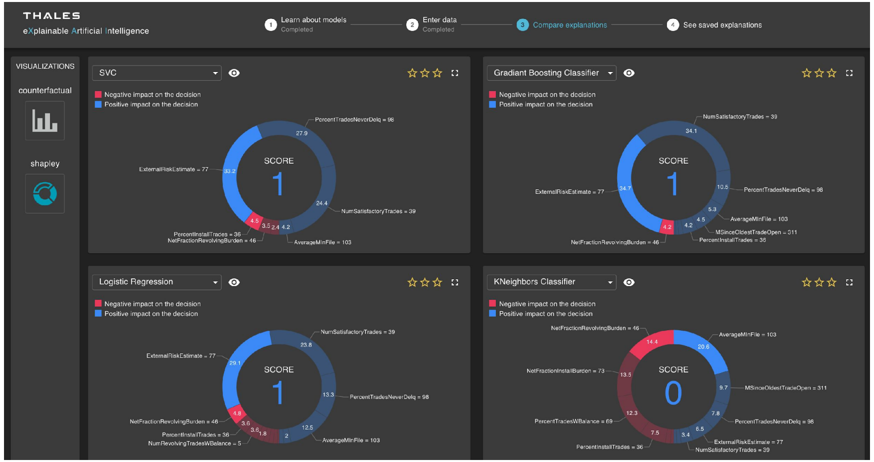
XAI Technology: Knowledge graphs and Artificial Neural Networks

THALES

Thales XAI Platform







Context

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

Goal

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

Approach: Model-Agnostic

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

THALES

https://xai.thalesdigitalsolutions.com/analysis

THALES
eXplainable Artificial Intelligence

1 Learn about models — Completed
2 Enter data — Completed
3 Compare explanations
4 See saved explanations

EXPLANATIONS

Lime

ResNet50 image classifier

Prediction: tank (n04389033) with proba: 0.8574951887130737

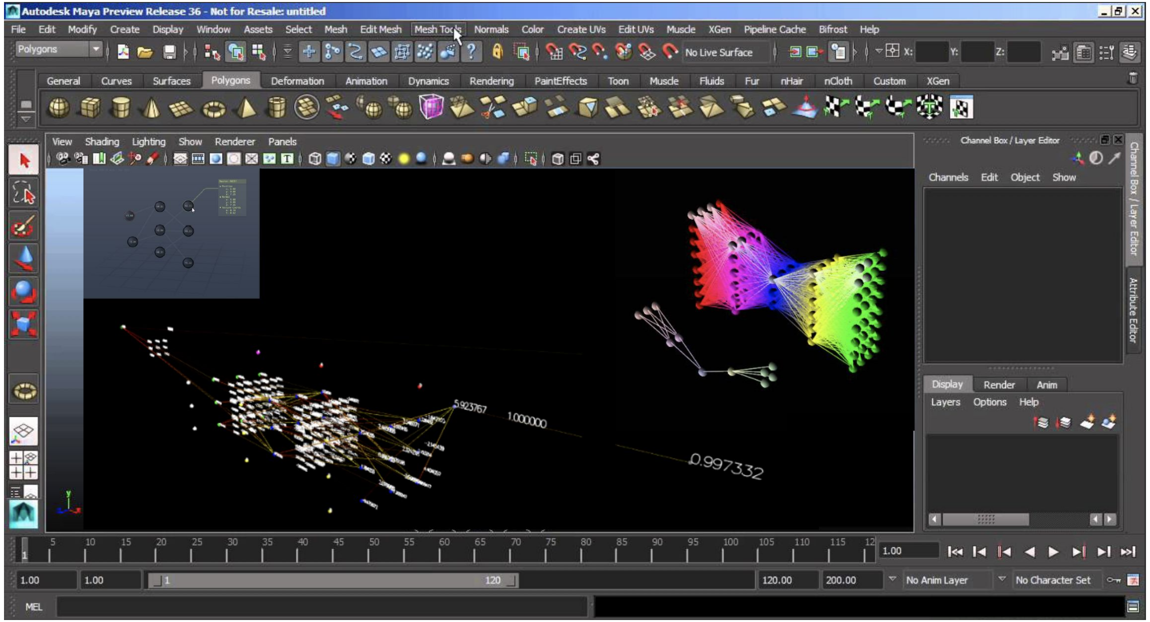
#1: Explaining Image Classification

Data: Image – XAI: Saliency Masks

BACK

NEXT

Debugging Artificial Neural Networks – Industry Agnostic

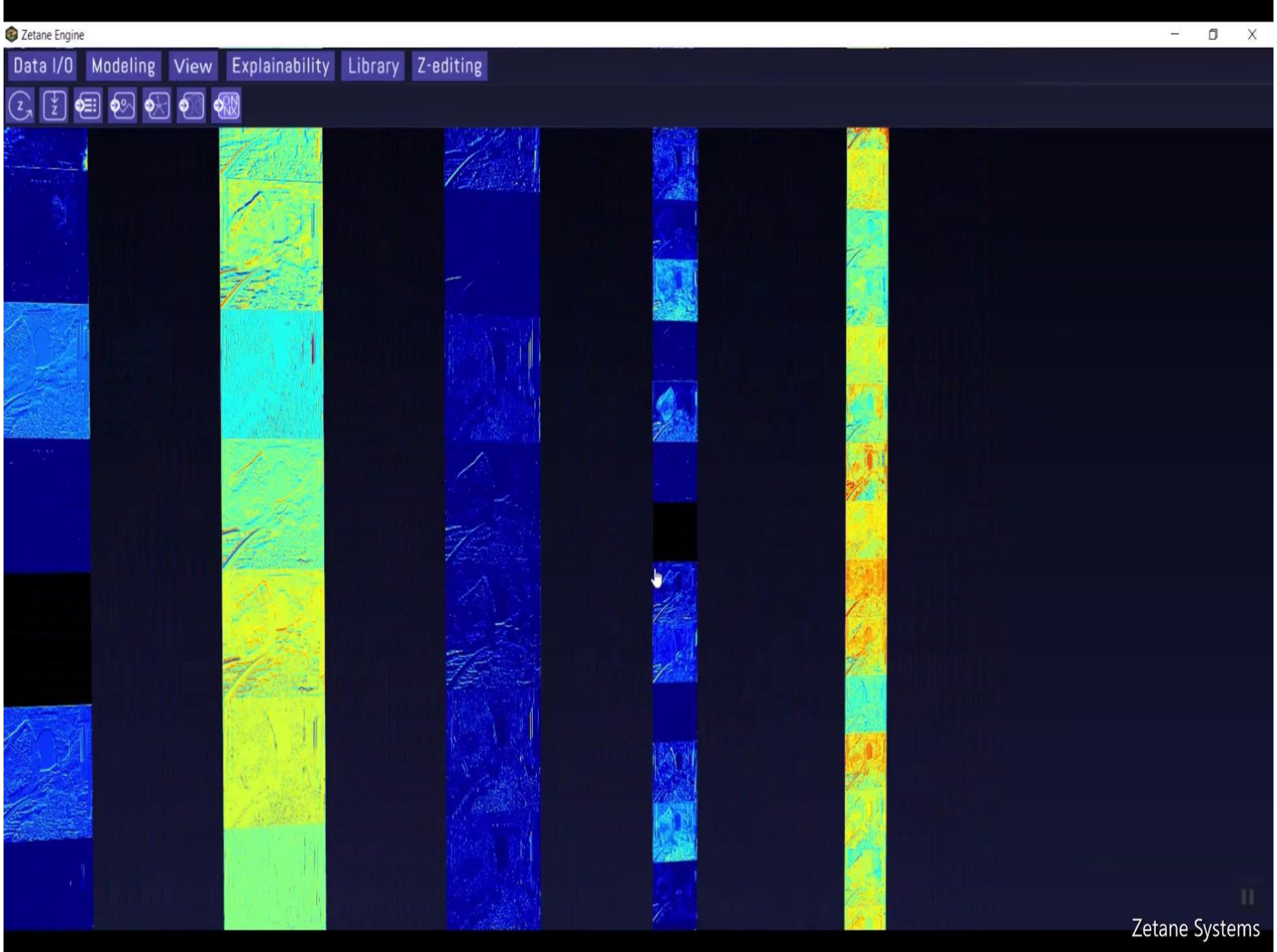


Challenge: Designing Artificial Neural Network architectures requires lots of experimentation (i.e., training phases) and parameters tuning (optimization strategy, learning rate, number of layers…) to reach optimal and robust machine learning models.

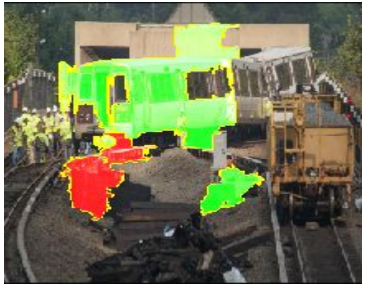
AI Technology: Artificial Neural Network

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

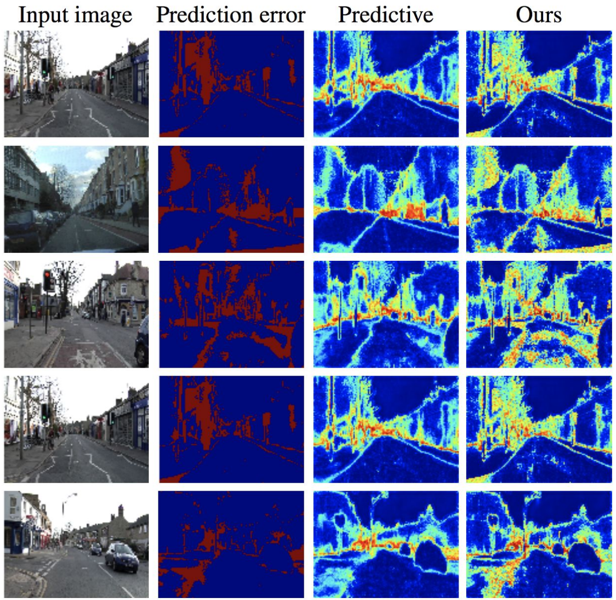
Zetane.com


Zetane Engine
Data I/O
Modeling
View
Explainability
Library
Z-editing
Zetane Systems

Obstacle Identification Certification (Trust) - Transportation



THALES



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

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

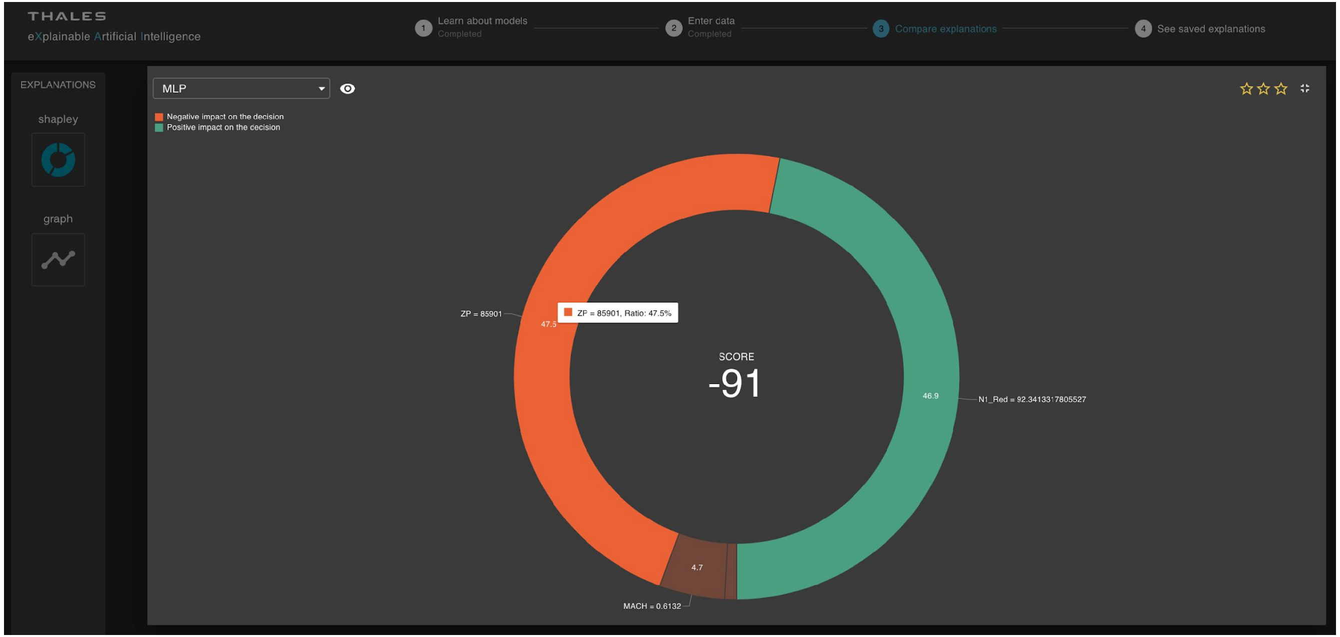
XAI Technology: Deep learning and Epistemic uncertainty

Explaining Flight Performance- Transportation

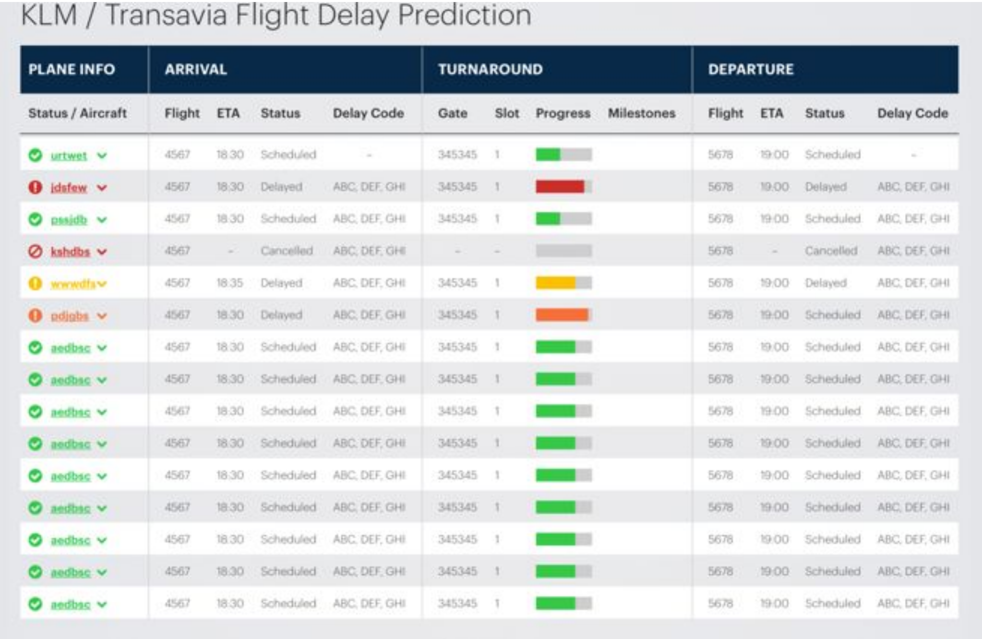
Challenge: Predicting and explaining aircraft engine performance

AI Technology: Artificial Neural Networks

XAI Technology: Shapely Values



Explainable On-Time Performance - Transportation



KLM / Transavia Flight Delay Prediction

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

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

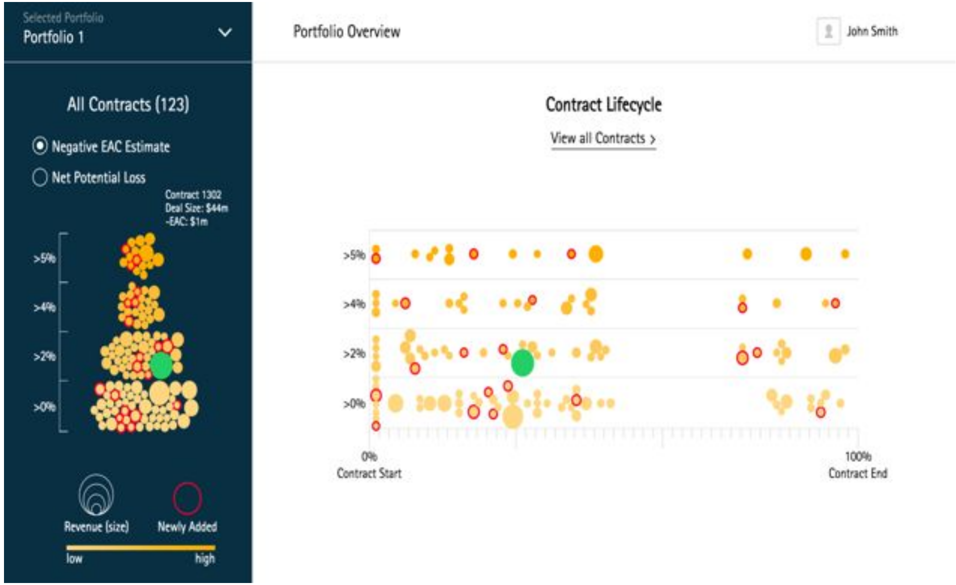
Nicholas McCarthy, Mohammad Karzand, Freddy Lecue: Amsterdam to Dublin Eventually Delayed? LSTM and Transfer Learning for Predicting Delays of Low Cost Airlines: AAAI 2019

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

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

XAI Technology: Knowledge graph embedded Sequence Learning using LSTMs

Explainable Risk Management - Finance



Jiewen Wu, Freddy Lécué, Christophe Guéret, Jer Hayes, Sara van de Moosdijk, Gemma Gallagher, Peter McCanney, Eugene Eichelberger: Personalizing Actions in Context for Risk Management Using Semantic Web Technologies. International Semantic Web Conference (2) 2017: 367-383

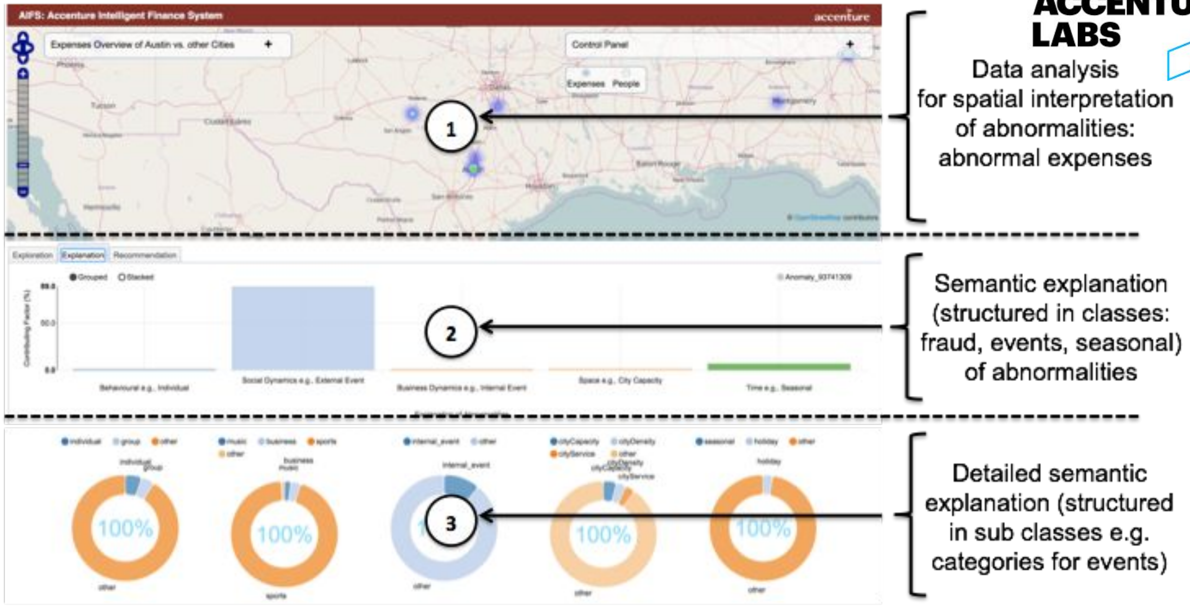
INNOVATION ARCHITECTURE: ACCENTURE LABS

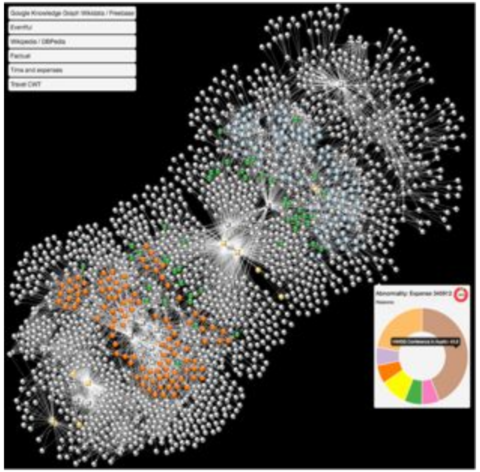
Challenge: Accenture is managing every year more than 80,000 opportunities and 35,000 contracts with an expected revenue of $34.1 billion. Revenue expectation does not meet estimation due to the complexity and risks of critical contracts. This is, in part, due to the (1) large volume of projects to assess and control, and (2) the existing non-systematic assessment process.

AI Technology: Integration of AI technologies i.e., Machine Learning, Reasoning, Natural Language Processing for building a robust model which can (1) predict revenue loss, (2) recommend corrective actions, and (3) explain why such actions might have a positive impact.

XAI Technology: Knowledge graph embedded Random Forrest

Explainable Anomaly Detection – Finance (Compliance)





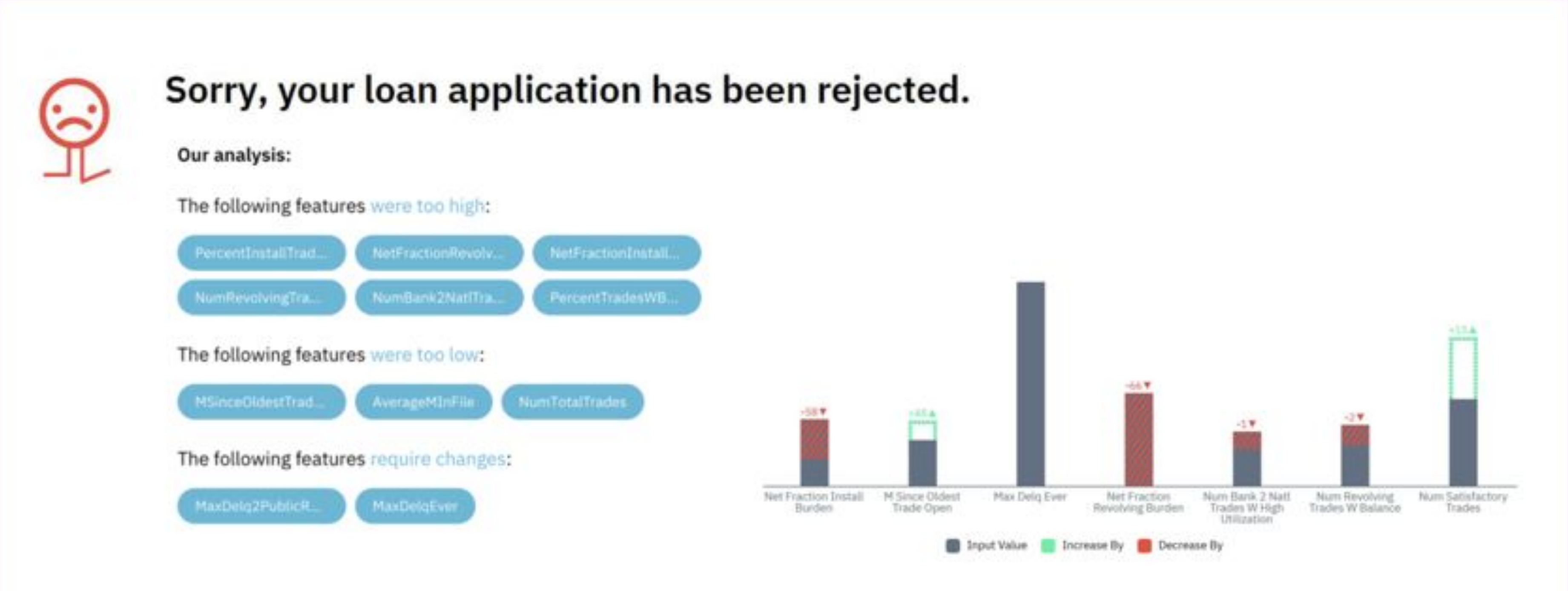
Freddy Lécué, Jiewen Wu: Explaining and predicting abnormal expenses at large scale using knowledge graph based reasoning. J. Web Sem. 44: 89-103 (2017)

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

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

XAI Technology: Knowledge graph embedded Ensemble Learning

Counterfactual Explanations for Credit Decisions (3) - Finance



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

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

Explanation of Medical Condition Relapse – Health

THALES

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

AI Technology: Relational learning

XAI Technology: Knowledge graphs and Artificial Neural Networks

Knowledge graph parts explaining medical condition relapse

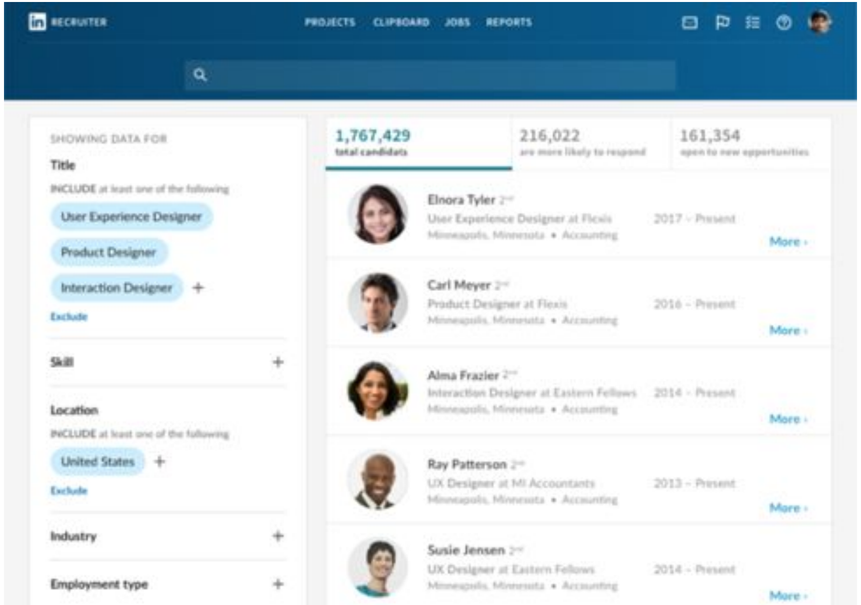
Case Study:

LinkedIn Talent Search

Varun Mithal, Girish Kathalagiri, Sahin Cem Geyik

LinkedIn Recruiter

- Recruiter Searches for Candidates
  - Standardized and free-text search criteria
- Retrieval and Ranking
  - Filter candidates using the criteria
  - Rank candidates in multiple levels using ML models

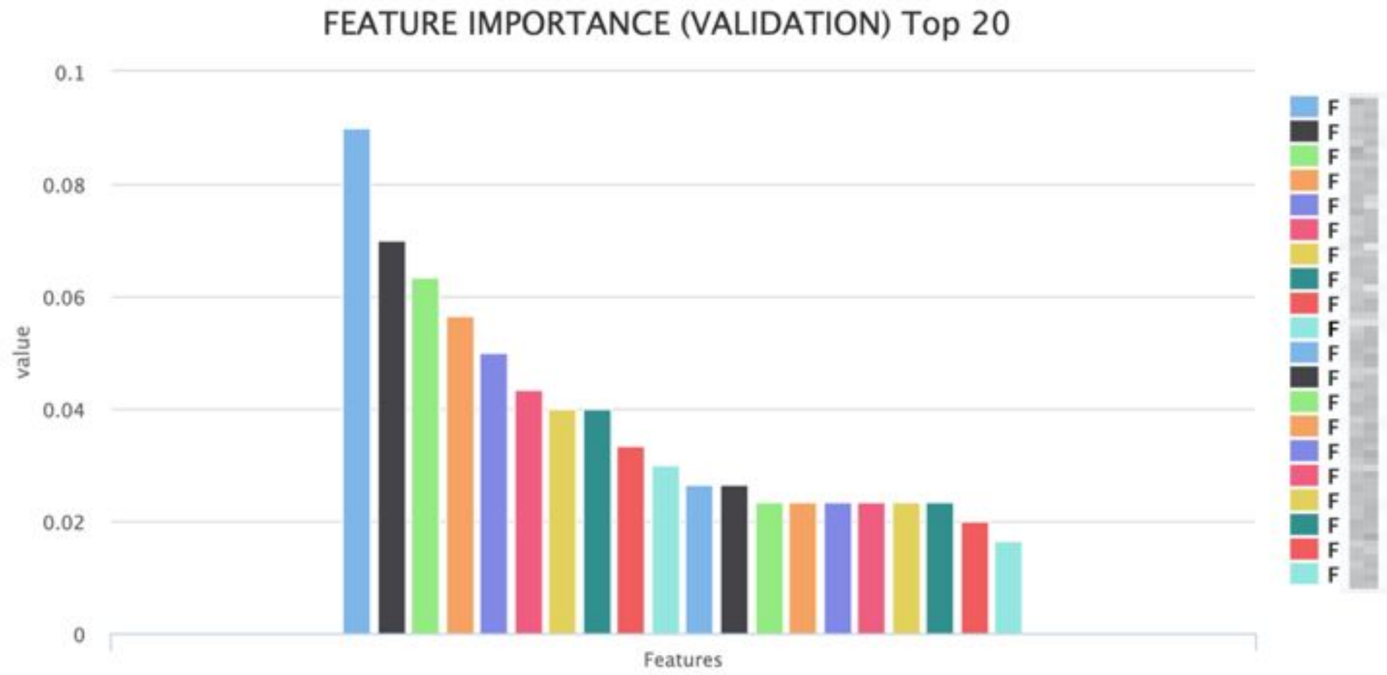


Modeling Approaches

- Pairwise XGBoost
- GLMix
- DNNs via TensorFlow

- Optimization Criteria: inMail Accepts
    - Positive: inMail sent by recruiter, and positively responded by candidate
        - Mutual interest between the recruiter and the candidate

Feature Importance in XGBoost



How We Utilize Feature Importances for GBDT

- Understanding feature digressions
  - Which a feature that was impactful no longer is?
  - Should we debug feature generation?
- Introducing new features in bulk and identifying effective ones
  - An activity feature for last 3 hours, 6 hours, 12 hours, 24 hours introduced (costly to compute)
  - Should we keep all such features?
- Separating the factors for that caused an improvement
  - Did an improvement come from a new feature, or a new labeling strategy, data source?
  - Did the ordering between features change?
- Shortcoming: A global view, not case by case

GLMix Models

- Generalized Linear Mixed Models
  - Global: Linear Model
  - Per-contract: Linear Model
  - Per-recruiter: Linear Model

$$g(\underbrace{P(r, c, re, ca, co)}_{\text{Positive Response Prob.}}) = \underbrace{\beta_{global} \cdot f_{all}}_{\text{Global model}} + \underbrace{\beta_{re} \cdot f_{all}}_{\text{Per-recruiter model}}$$
$$+ \underbrace{\beta_{co} \cdot f_{all}}_{\text{Per-contract model}}$$

- Lots of parameters overall
  - For a specific recruiter or contract the weights can be summed up
- Inherently explainable
  - Contribution of a feature is “weight x feature value”
  - Can be examined in a case-by-case manner as well

TensorFlow Models in Recruiter and Explaining Them

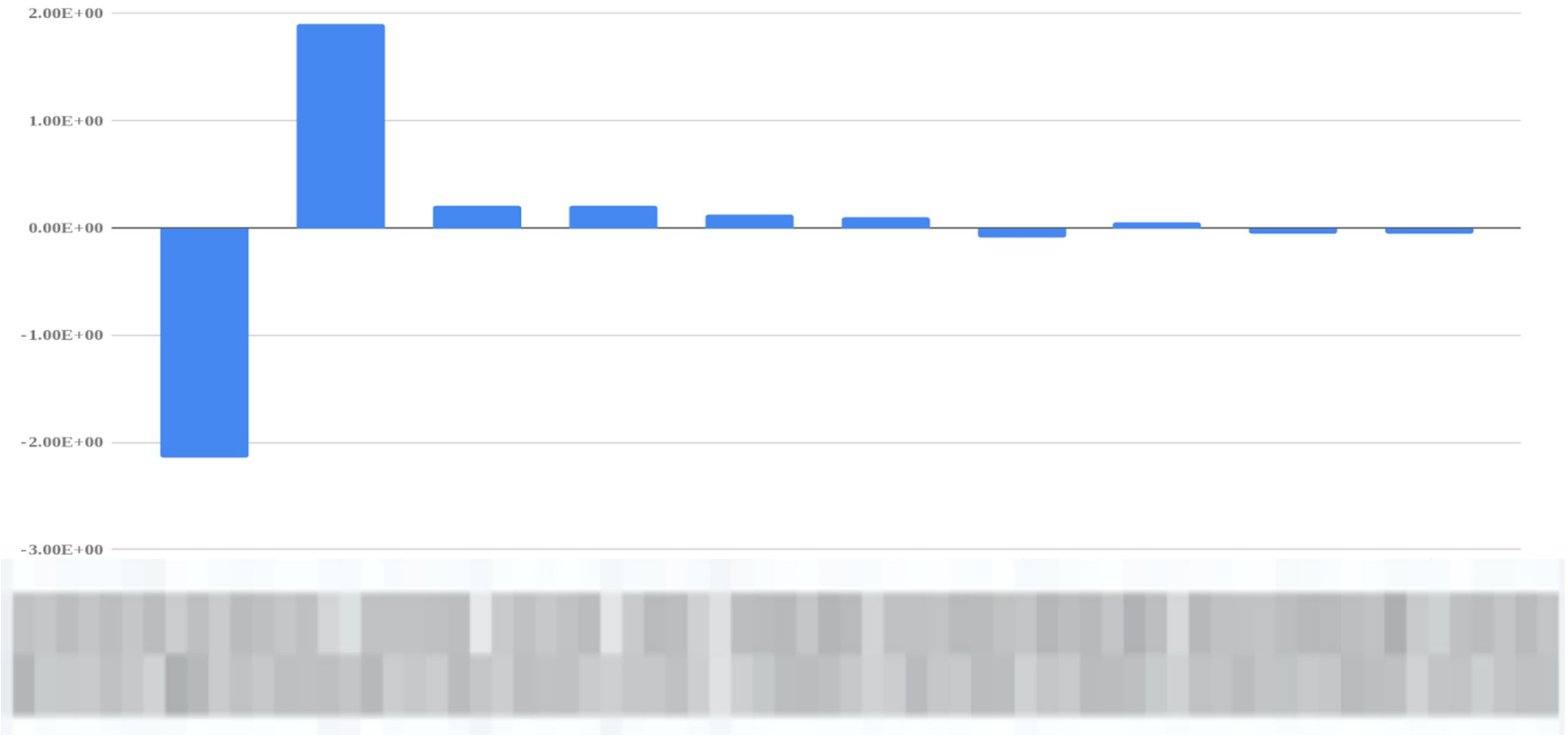
- We utilize the Integrated Gradients [ICML 2017] method

- How do we determine the baseline example?
  - Every query creates its own feature values for the same candidate
  - Query match features, time-based features
  - Recruiter affinity, and candidate affinity features
  - A candidate would be scored differently by each query
  - Cannot recommend a “Software Engineer” to a search for a “Forensic Chemist”
  - There is no globally neutral example for comparison!

Query-Specific Baseline Selection

- For each query:
  - Score examples by the TF model
  - Rank examples
  - Choose one example as the baseline
  - Compare others to the baseline example
- How to choose the baseline example
  - Last candidate
  - Kth percentile in ranking
  - A random candidate
  - Request by user (answering a question like: “Why was I presented candidate x above candidate y?”)

Example



Example - Detailed

Feature	**Description**	**Difference (1 vs 2)**	**Contribution**
Feature……….	Description……….	-2.0476928	-2.144455602
Feature……….	Description……….	-2.3223877	1.903594618
Feature……….	Description……….	0.11666667	0.2114946752
Feature……….	Description……….	-2.1442587	0.2060414469
Feature……….	Description……….	-14	0.1215354111
Feature……….	Description……….	1	0.1000282466
Feature……….	Description……….	-92	-0.085286277
Feature……….	Description……….	0.9333333	0.0568533262
Feature……….	Description……….	-1	-0.051796317
Feature……….	Description……….	-1	-0.050895940

Pros & Cons

- Explains potentially very complex models
- Case-by-case analysis
  - Why do you think candidate x is a better match for my position?
  - Why do you think I am a better fit for this job?
  - Why am I being shown this ad?
  - Great for debugging real-time problems in production
- Global view is missing
  - Aggregate Contributions can be computed
  - Could be costly to compute

Lessons Learned and Next Steps

- Global explanations vs. Case-by-case Explanations
  - Global gives an overview, better for making modeling decisions
  - Case-by-case could be more useful for the non-technical user, better for debugging
- Integrated gradients worked well for us
  - Complex models make it harder for developers to map improvement to effort
  - Use-case gave intuitive results, on top of completely describing score differences
- Next steps
  - Global explanations for Deep Models

Case Study:

Model Interpretation for Predictive Models in B2B Sales Predictions

Jilei Yang, Wei Di, Songtao Guo

LinkedIn

Problem Setting

- Predictive models in B2B sales prediction
  - E.g.: random forest, gradient boosting, deep neural network, …
  - High accuracy, low interpretability
- Global feature importance → Individual feature reasoning

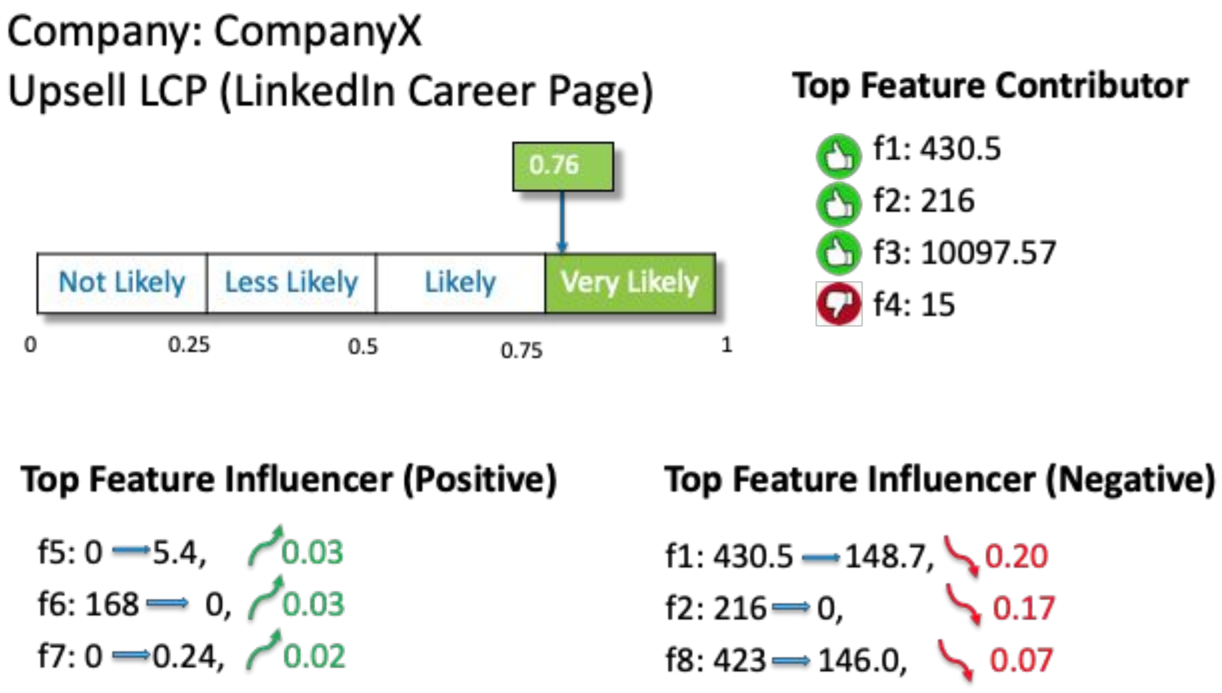
① What are top driver features for a certain company to have high/low probability to upsell/churn?

① Feature Contributor

② Which top driver features can be perturbed if we want to increase/decrease probability for a certain company?

② Feature Influencer

Example


Company: CompanyX
Upsell LCP (LinkedIn Career Page)
0.76
Not Likely
Less Likely
Likely
Very Likely
0
0.25
0.5
0.75
1
Top Feature Contributor
f1: 430.5
f2: 216
f3: 10097.57
f4: 15
Top Feature Influencer (Positive)
f5: 0 → 5.4, 0.03
f6: 168 → 0, 0.03
f7: 0 → 0.24, 0.02
Top Feature Influencer (Negative)
f1: 430.5 → 148.7, 0.20
f2: 216 → 0, 0.17
f8: 423 → 146.0, 0.07

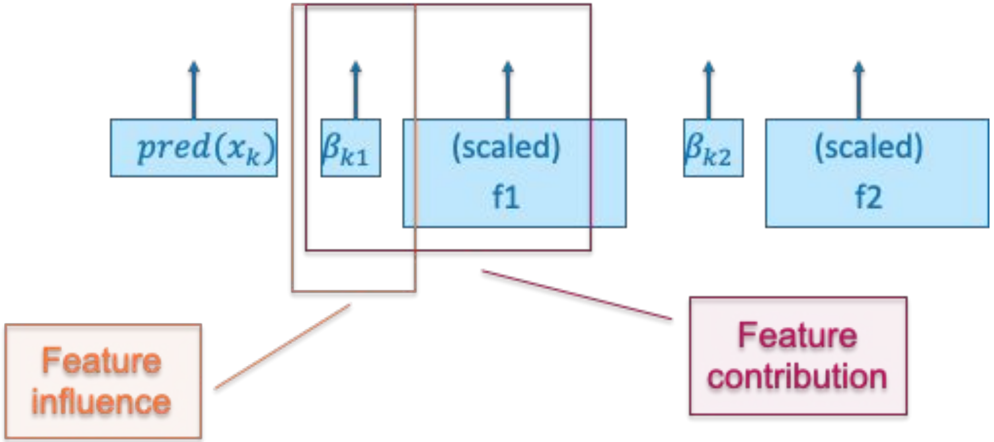
Revisiting LIME

- Given a target sample x_k, approximate its prediction $pred(x_k)$ by building a sample-specific linear model:

$$pred(X) \approx \beta_{k1} X_1 + \beta_{k2} X_2 + \ldots, X \in neighbor(x_k)$$

- E.g., for company CompanyX:

$$0.76 \approx 1.82 * 0.17 \quad + 1.61 * 0.11 + \ldots$$



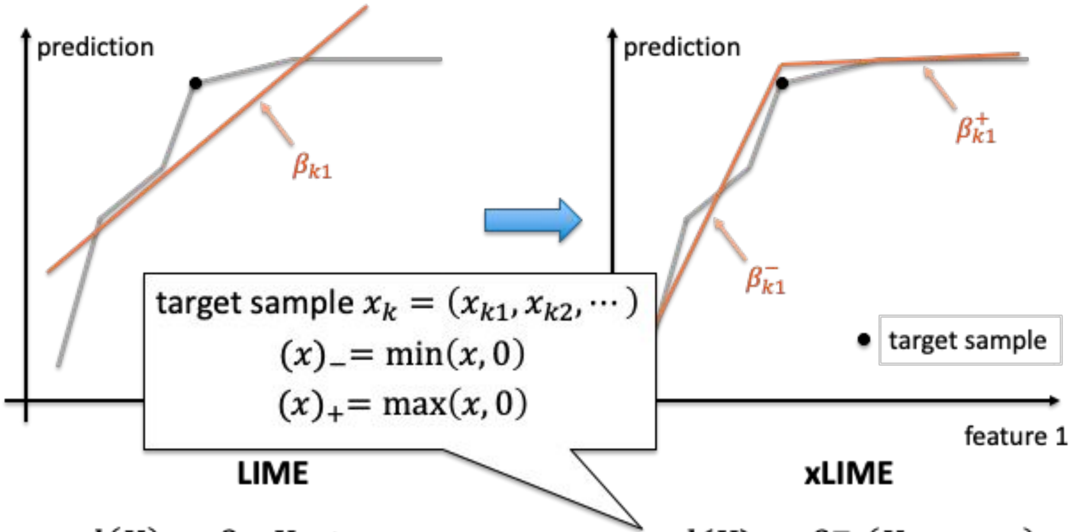
xLIME

Piecewise Linear Regression

Localized Stratified Sampling

Piecewise Linear Regression

Motivation: Separate top positive feature influencers and top negative feature influencers



$$pred(X) \approx \beta_{k1} X_1 + \cdots,$$
$$X \in neighbor(x_k)$$

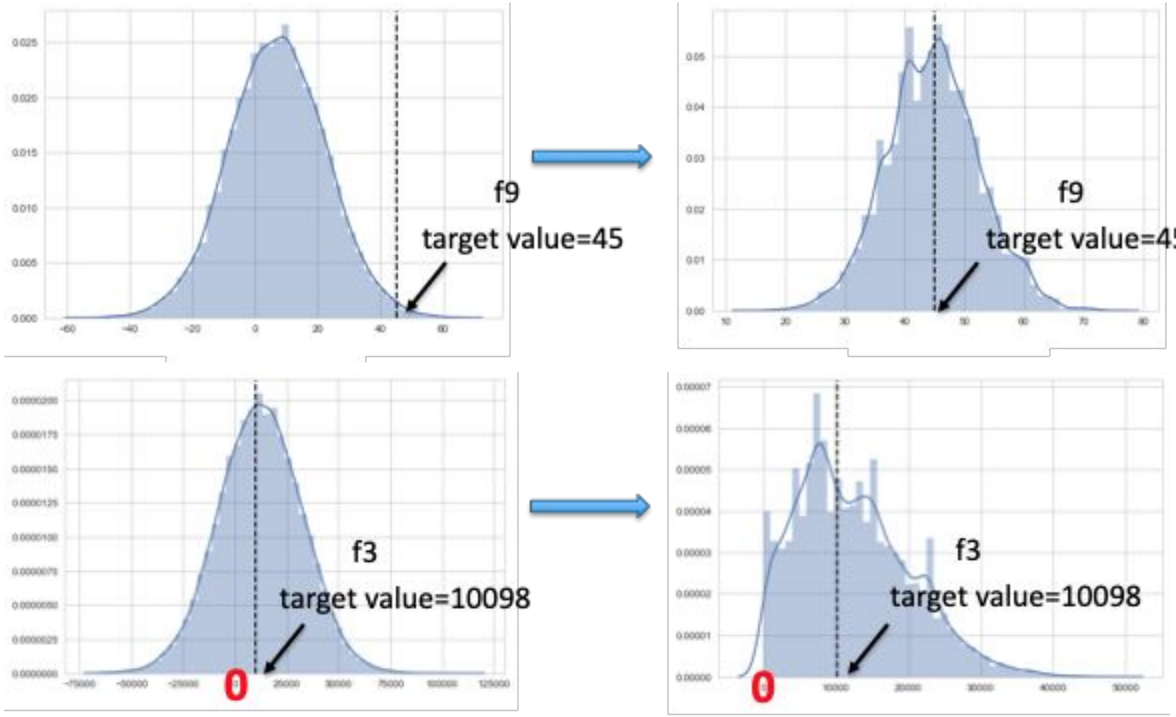
$$pred(X) \approx \beta_{k1}^{-}(X_1 - x_{k1})_-$$
$$+\beta_{k1}^{+}(X_1 - x_{k1})_+ + \cdots,$$
$$X \in neighbor(x_k)$$

Impact of Piecewise Approach

- Target sample x_k=(x_{k1}, x_{k2}, ⋯)
- Top feature contributor
  - LIME: large magnitude of $\beta_{kj} \cdot x_{kj}$
  - xLIME: large magnitude of $\beta_{kj}^{-} \cdot x_{kj}$
- Top positive feature influencer
  - LIME: large magnitude of β_{kj}
  - xLIME: large magnitude of negative β_{kj}^{-} or positive β_{kj}^{+}
- Top negative feature influencer
  - LIME: large magnitude of β_{kj}
  - xLIME: large magnitude of positive β_{kj}^{-} or negative β_{kj}^{+}

Localized Stratified Sampling: Idea

Method: Sampling based on empirical distribution around target value at each feature level



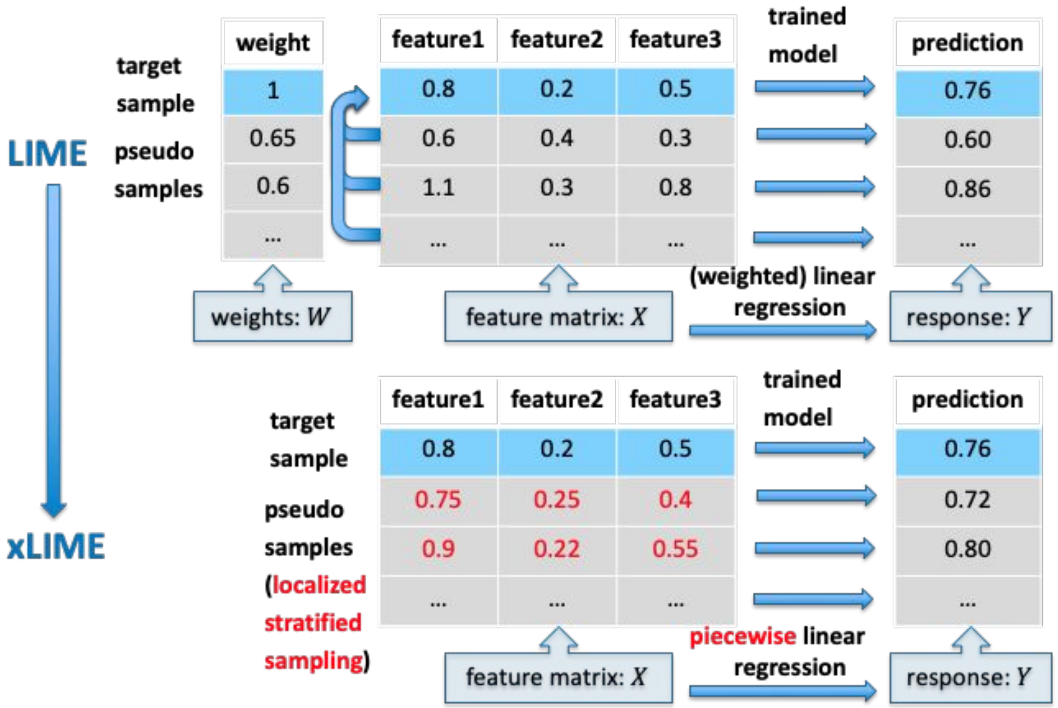
Localized Stratified Sampling: Method

- Sampling based on empirical distribution around target value for each feature
- For target sample $x_k = (x_{k1}, x_{k2}, \cdots)$, sampling values of feature j according to

$$p_j(X_j) \cdot N(x_{kj}, (\alpha \cdot s_j)^2)$$

  - $p_j(X_j)$: empirical distribution.
  - x_{kj} : feature value in target sample.
  - s_j : standard deviation.
  - α : Interpretable range: tradeoff between interpretable coverage and local accuracy.
- In LIME, sampling according to $N(\bar{x}_j, s_j^2)$.

Summary


LIME
xLIME
target sample
pseudo samples
weight
1
0.65
0.6
...
feature1
feature2
feature3
0.8
0.2
0.5
0.6
0.4
0.3
1.1
0.3
0.8
trained model
prediction
0.76
0.60
0.86
weights: W
feature matrix: X
(weighted) linear regression
response: Y
target sample
pseudo samples (localized stratified sampling)
feature1
feature2
feature3
0.8
0.2
0.5
0.75
0.25
0.4
0.9
0.22
0.55
trained model
prediction
0.76
0.72
0.80
feature matrix: X
piecewise linear regression
response: Y

LTS LCP (LinkedIn Career Page) Upsell

- A subset of churn data
  - Total Companies: ~ 19K
  - Company features: 117
- **Problem:** Estimate whether there will be upsell given a set of features about the company's utility from the product

Top Feature Contributor

Company : CompanyX

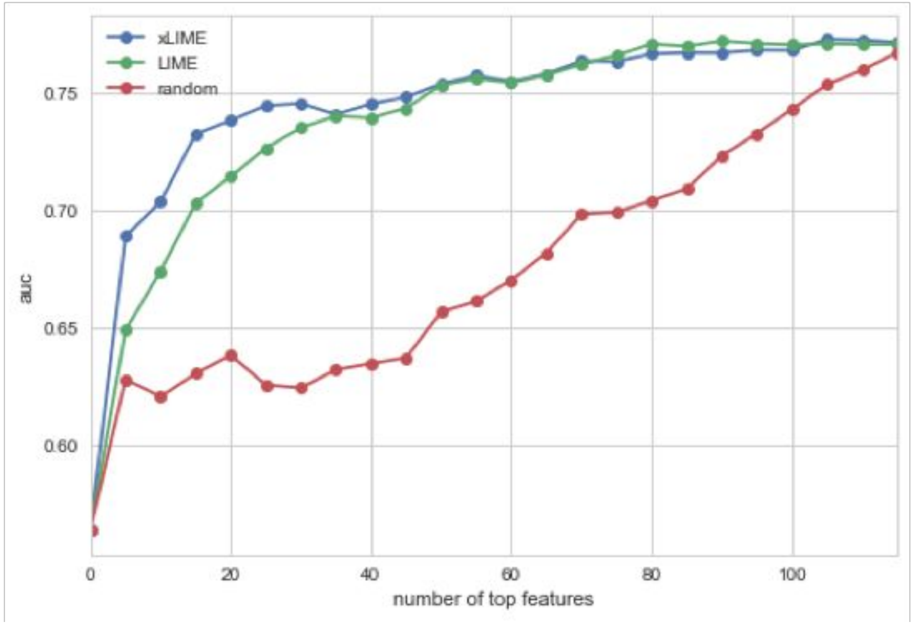
LIME

	name	value	quantile	contribution
	f9	45.0	98	-0.011
	f3	10097.6	66	0.011
	f10	16.5	94	0.010

xLIME

	name	value	quantile	contribution
	f1	430.5	59	0.246
	f2	216.0	40	0.161
	f3	10097.6	66	0.084

- **Explanation curve**: how classification performance varies if one considers only the top ranked feature contributors



Top Feature Influencers

Company: CompanyX

	Positive influencer			Negative influencer		
LIME	f1 +	430.5→712.3	.004	f1 -	430.5→148.7	.004
	f2 +	216.0→435.4	.004	f2 -	216.0→0.0	.004
	f11 +	9.8→13.2	.003	f11 -	9.8→6.3	.003
xLIME	f5 +	0.0→5.4	.032	f1 -	430.5→148.7	.201
	f6 -	168.0→0.0	.031	f2 -	216.0→0.0	.174
	f7 +	0.00→0.24	.016	f8 -	423.0→146.0	.071

Key Takeaways

- Looking at the explanation as contributor vs. influencer features is useful
  - Contributor: Which features end-up in the current outcome case-by-case
  - Influencer: **What needs to be done to improve likelihood, case-by-case**

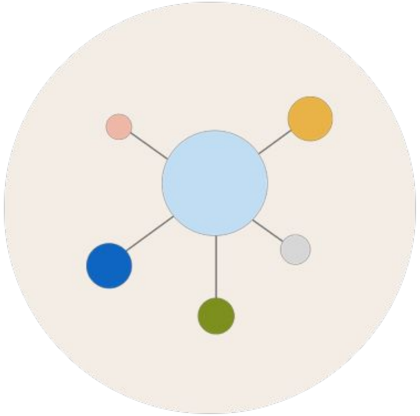
- xLIME aims to improve on LIME via:
  - Piecewise linear regression: More accurately describes local point, helps with finding correct influencers
  - Localized stratified sampling: More realistic set of local points
- Better captures the important features

Case Study:

Relevance Debugging and Explaining @ LinkedIn

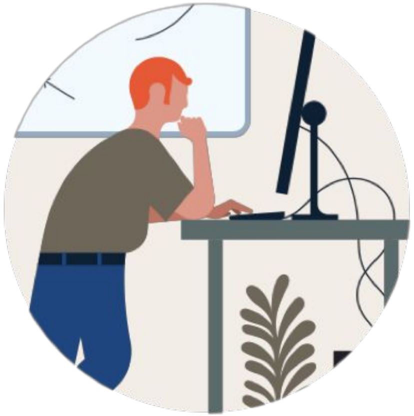
Daniel Qiu, Yucheng Qian

Debugging Relevance Models



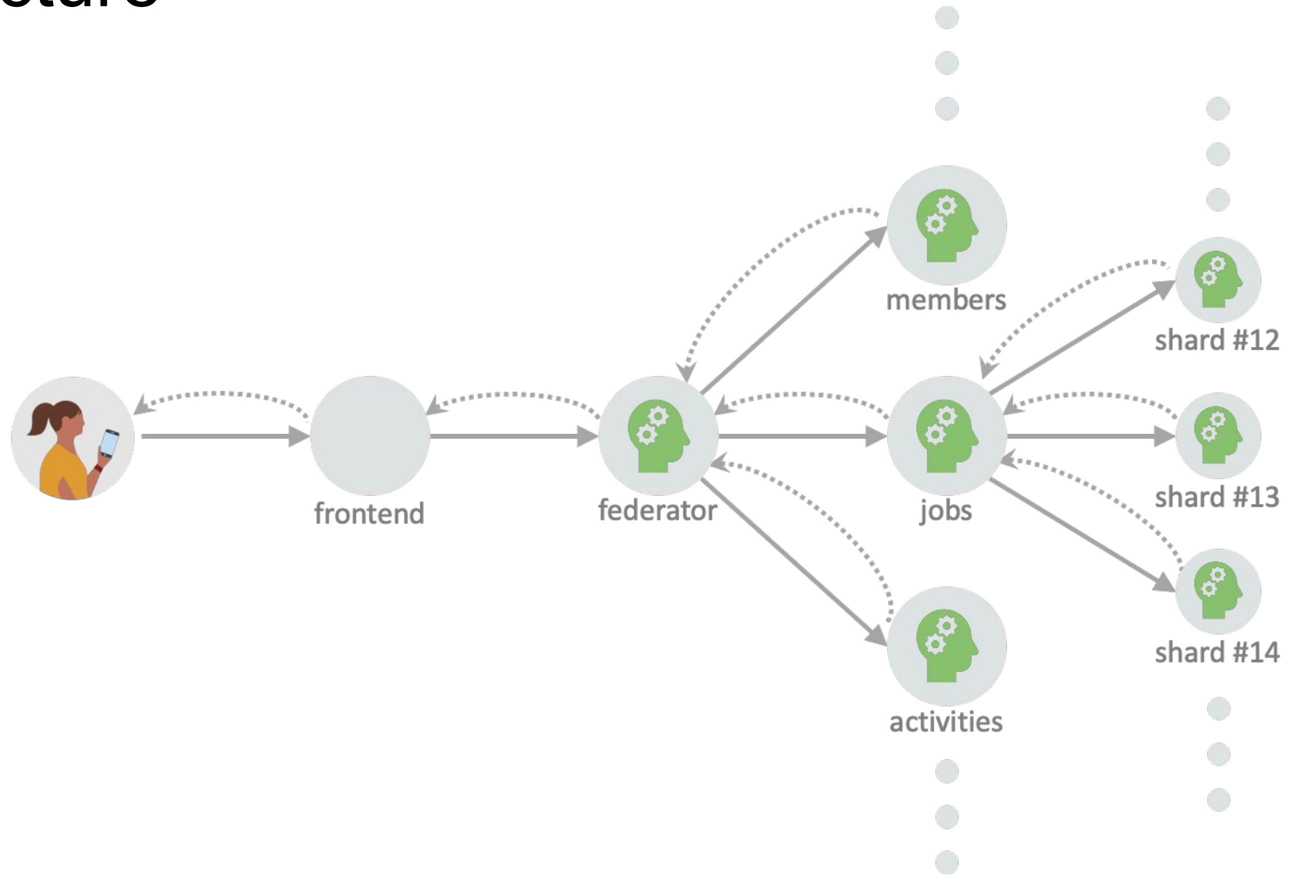
Modeling
Improve the machine learning model

Value
Bring value to our members by providing relevant experience

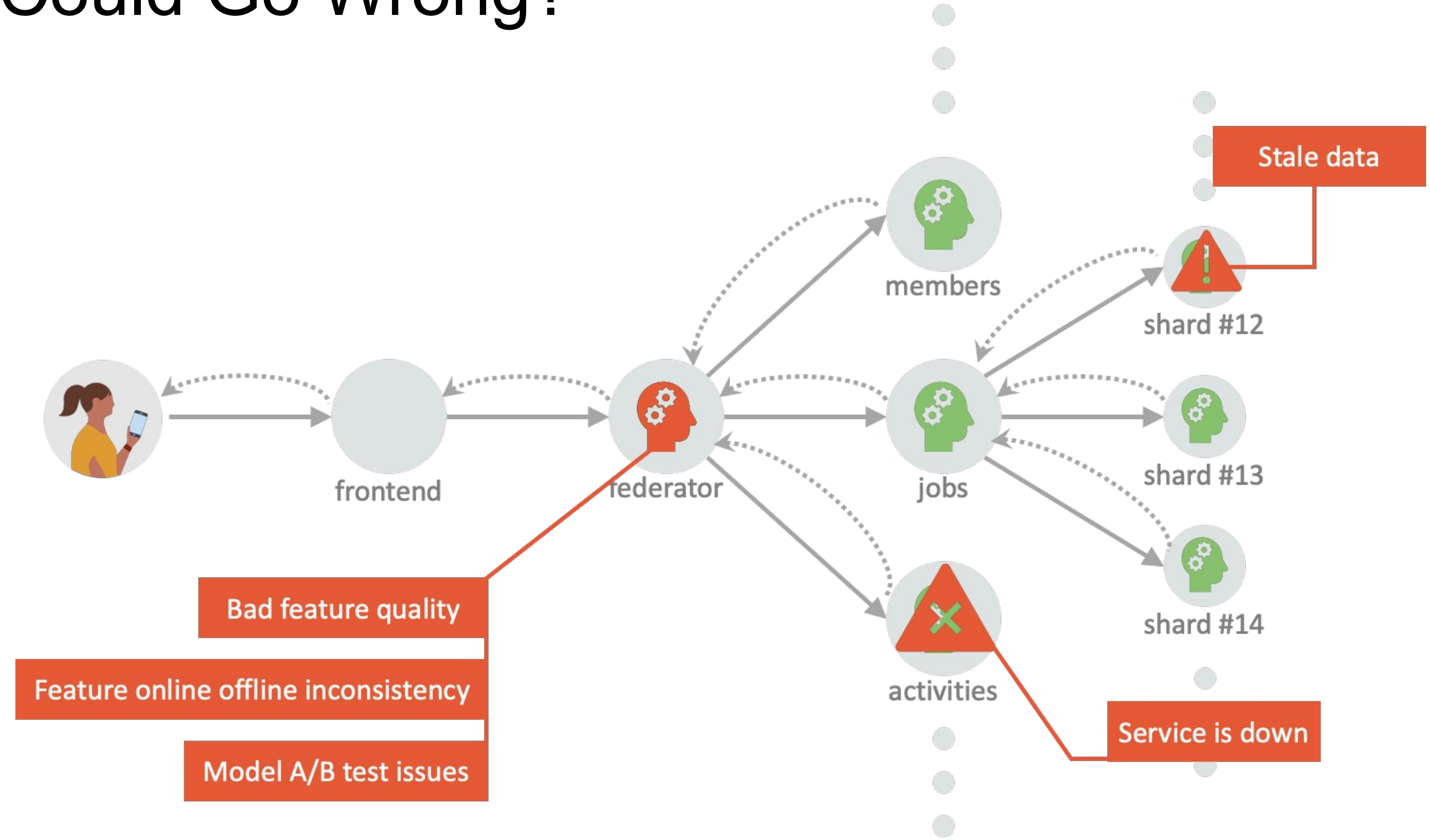


Trust
Build trust with our members

Architecture



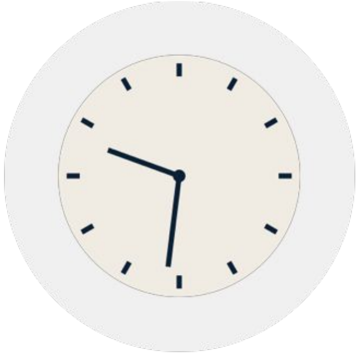
What Could Go Wrong?



Challenges

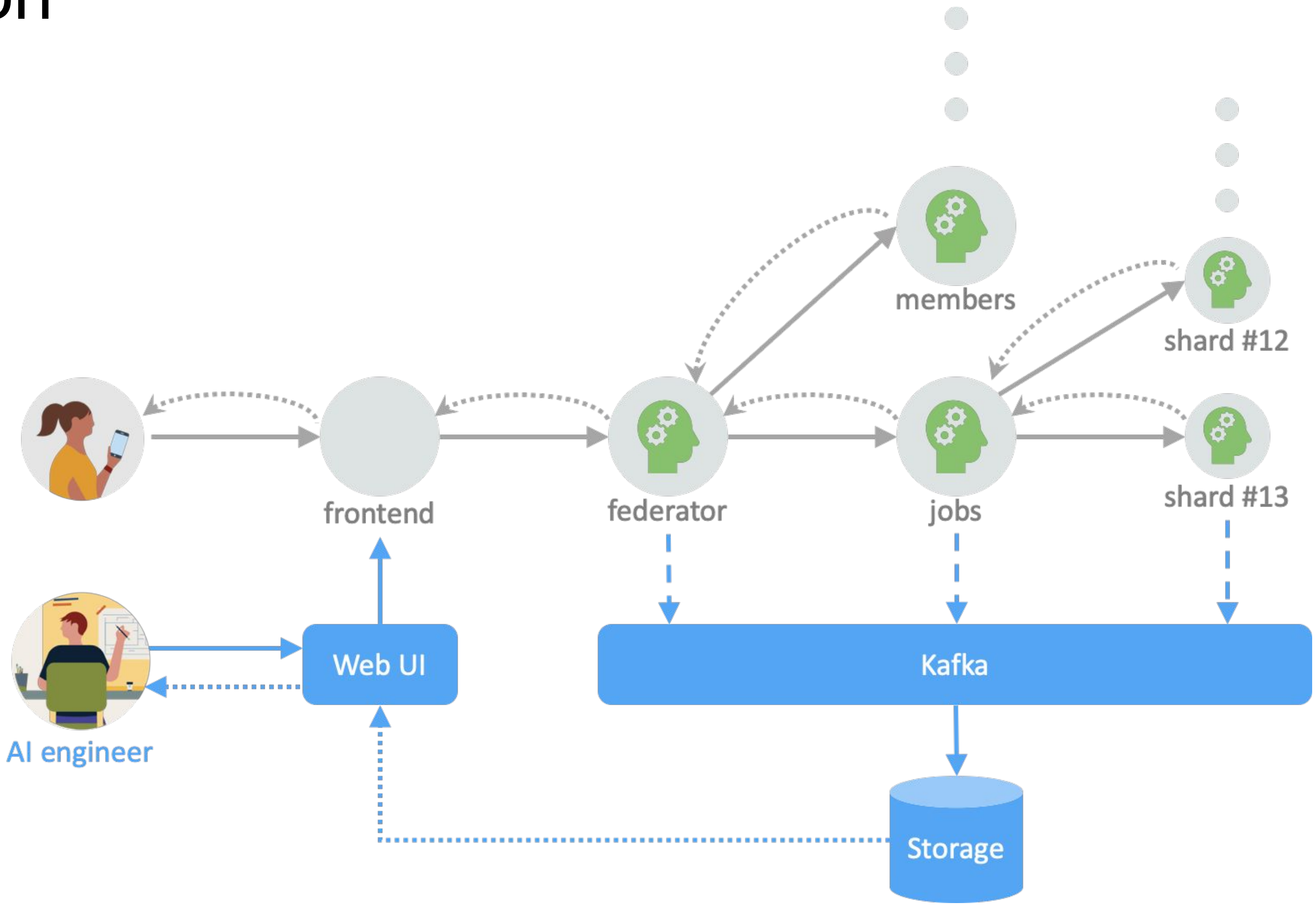
Complex Infrastructure

Hard to Reproduce



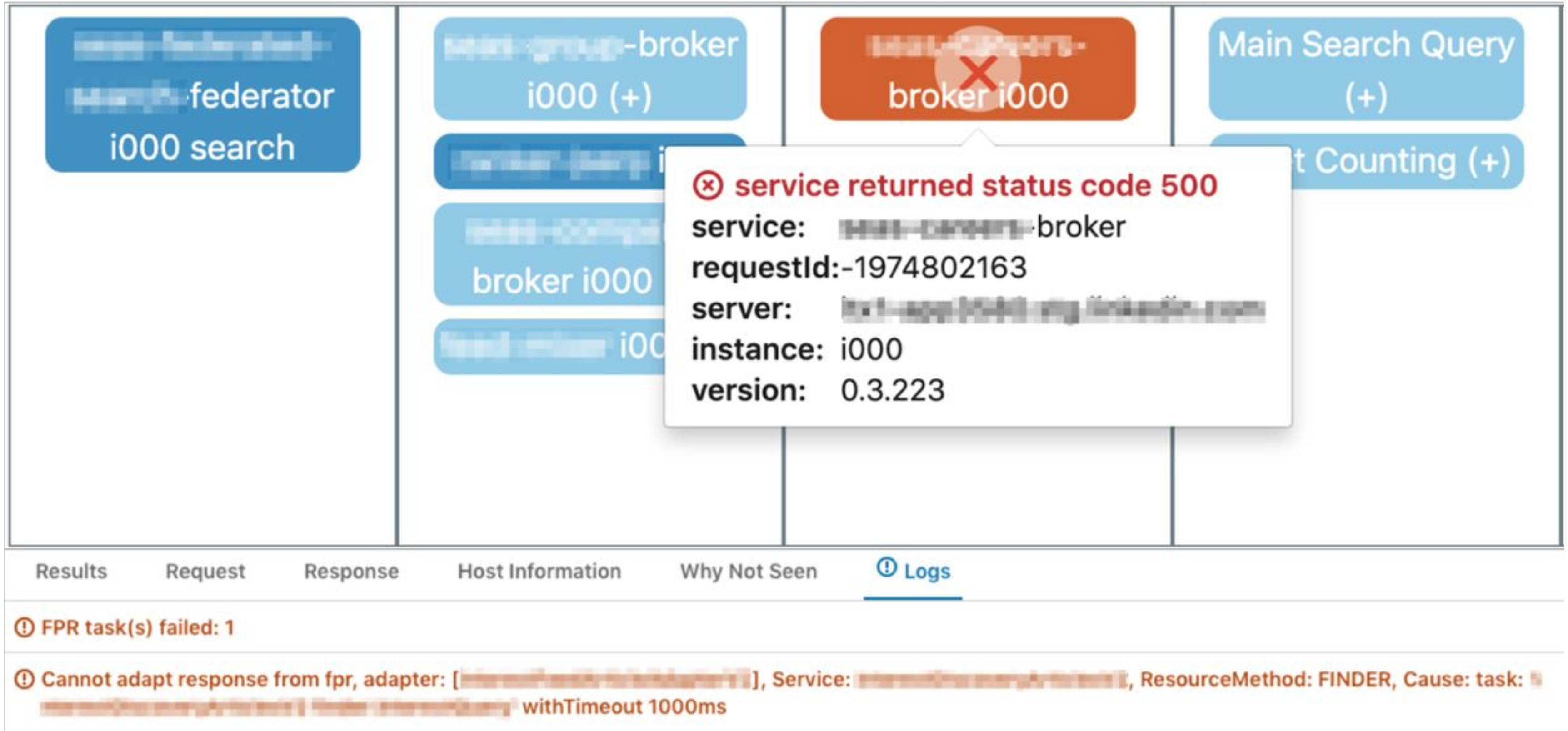
Time Consuming

Solution



members
shard #12
frontend
federator
jobs
shard #13
Web UI
Kafka
AI engineer
Storage

Call Graph



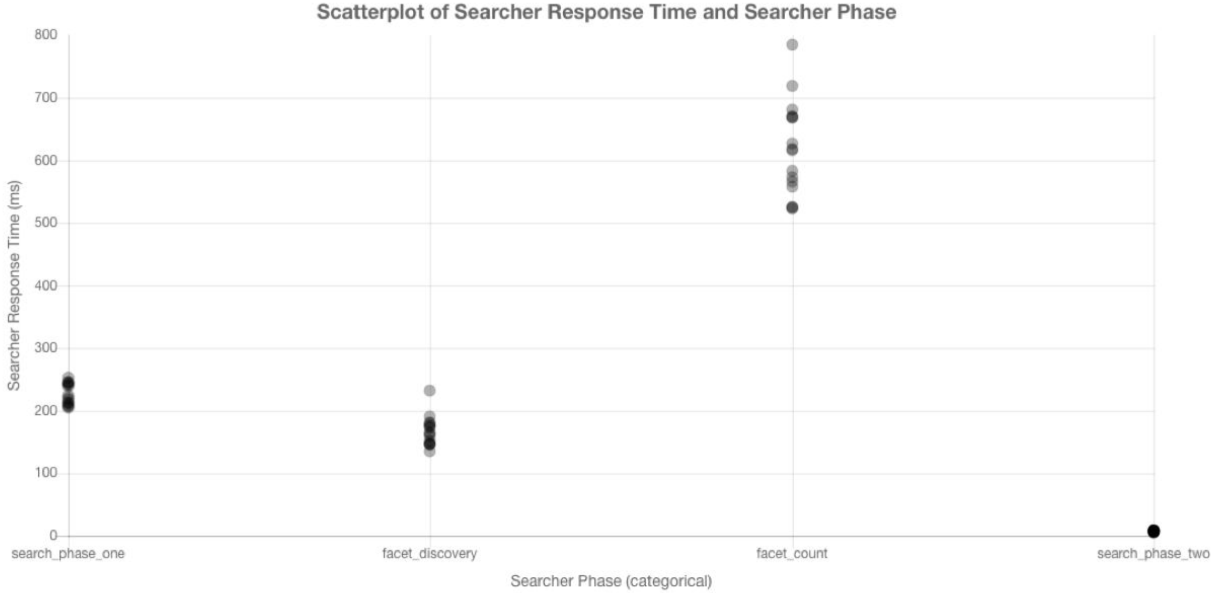
federator
i000 search
-broker
i000 (+)
broker i000
broker i000
Main Search Query
(+)
t Counting (+)
service returned status code 500
service: broker
requestId:-1974802163
server:
instance: i000
version: 0.3.223
Results
Request
Response
Host Information
Why Not Seen
Logs
FPR task(s) failed: 1
Cannot adapt response from fpr, adapter: [ ], Service: , ResourceMethod: FINDER, Cause: task:
withTimeout 1000ms

Timing

Total time (ms): 1041

Number of garbage collection events: 0

	Start Time	End Time	Total Time	Resent?	Partitions	Min	Max	p50	p90
search_phase_one	7	266	259	false	16	205	253	223.0	245.5
facet_discovery	13	240	227	true	16	135	232	164.0	186.0
facet_count	262	1041	779	true	16	523	785	617.0	700.0
search_phase_two	266	274	8	false	15	5	9	8.0	9.0

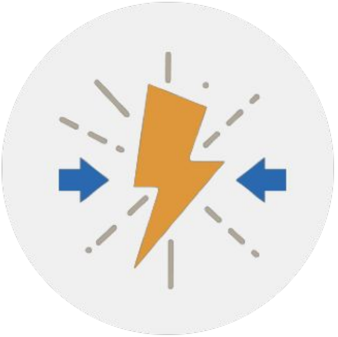


Features

Group	Feature	Value
SPR	activity_recent_click /	968
SPR	[illegible]	1
SPR	[illegible]	6.8762646
SPR	[illegible]	null
SPR	[illegible]	null
SPR	binary_activity_recent_click / |	1
SPR	[illegible]	null
SPR	log_activity_recent_click /	6.8762646
SPR	[illegible]	0
SPR	[illegible]	0

Advanced Use Cases

Perturbation



Comparison

Replay

Perturbation

1. Inject

Injected as part of the request

- Override A/B test settings
- Model selection
- Feature override

2. Relay

Passed to downstream service

3. Overwrite

Overwrite the system behavior

Comparison

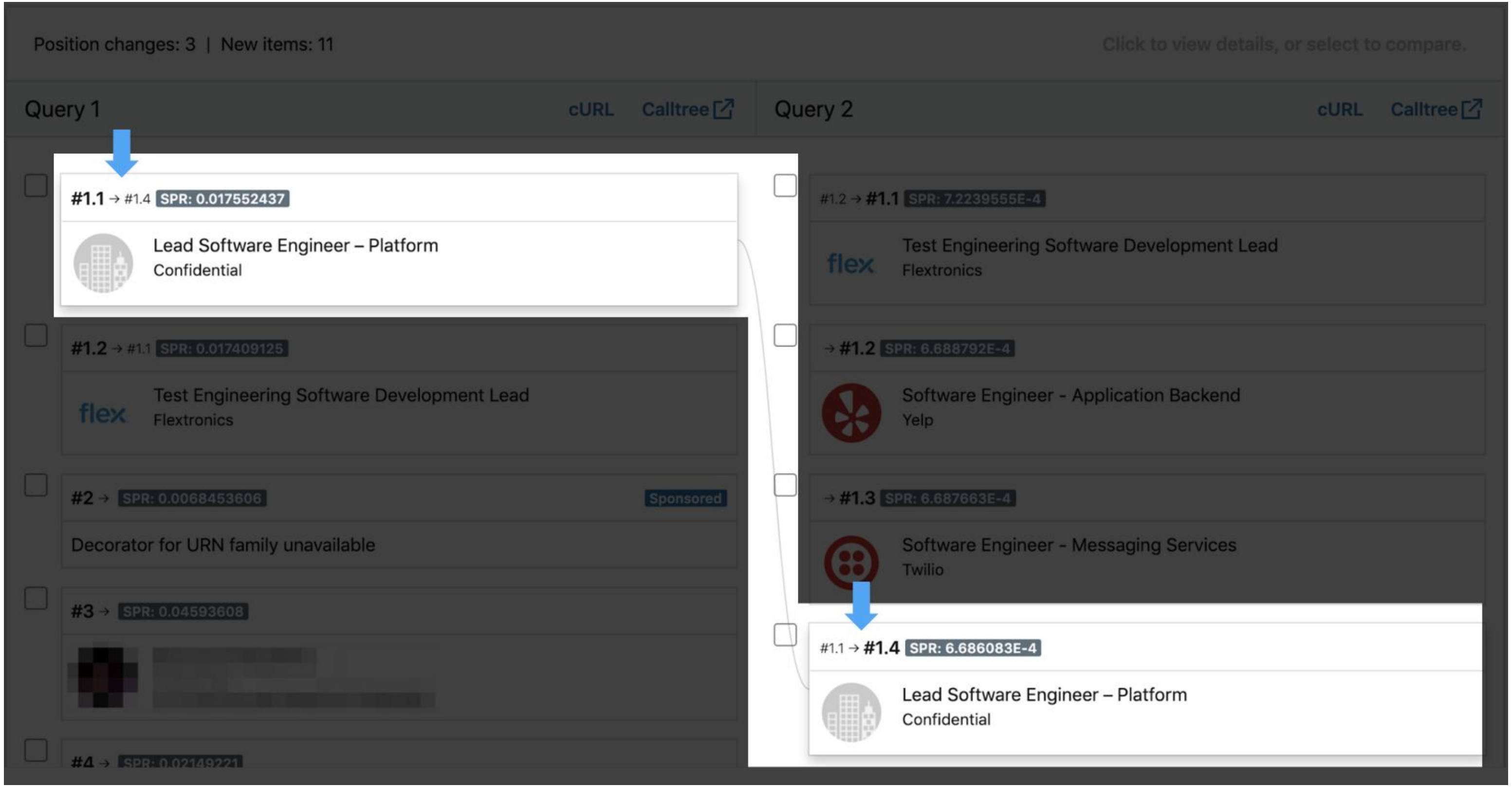
Compare Model

Compare results of 2 different queries/models

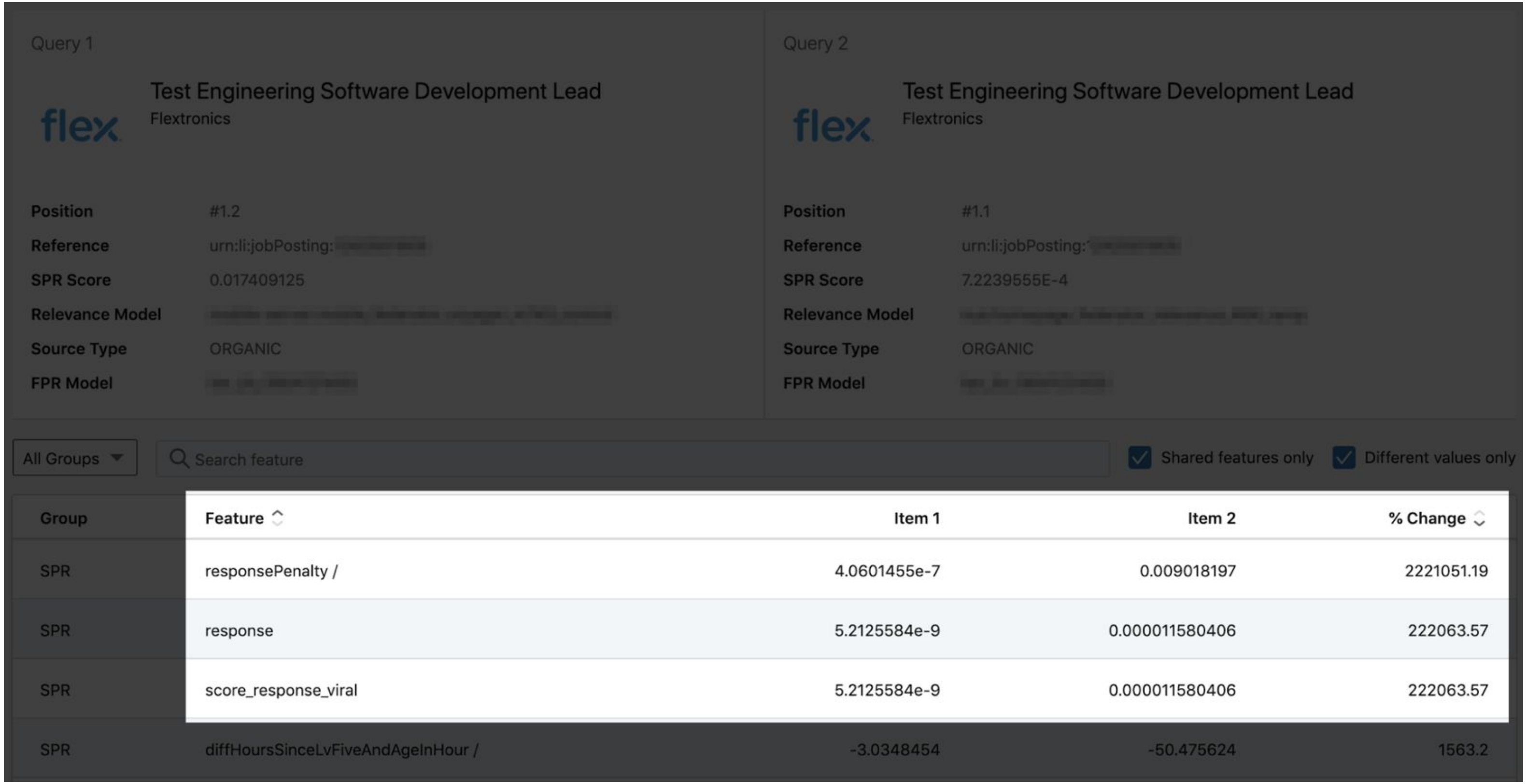
Compare Items

Compare features and scores of 2 different items, from the same query or different queries

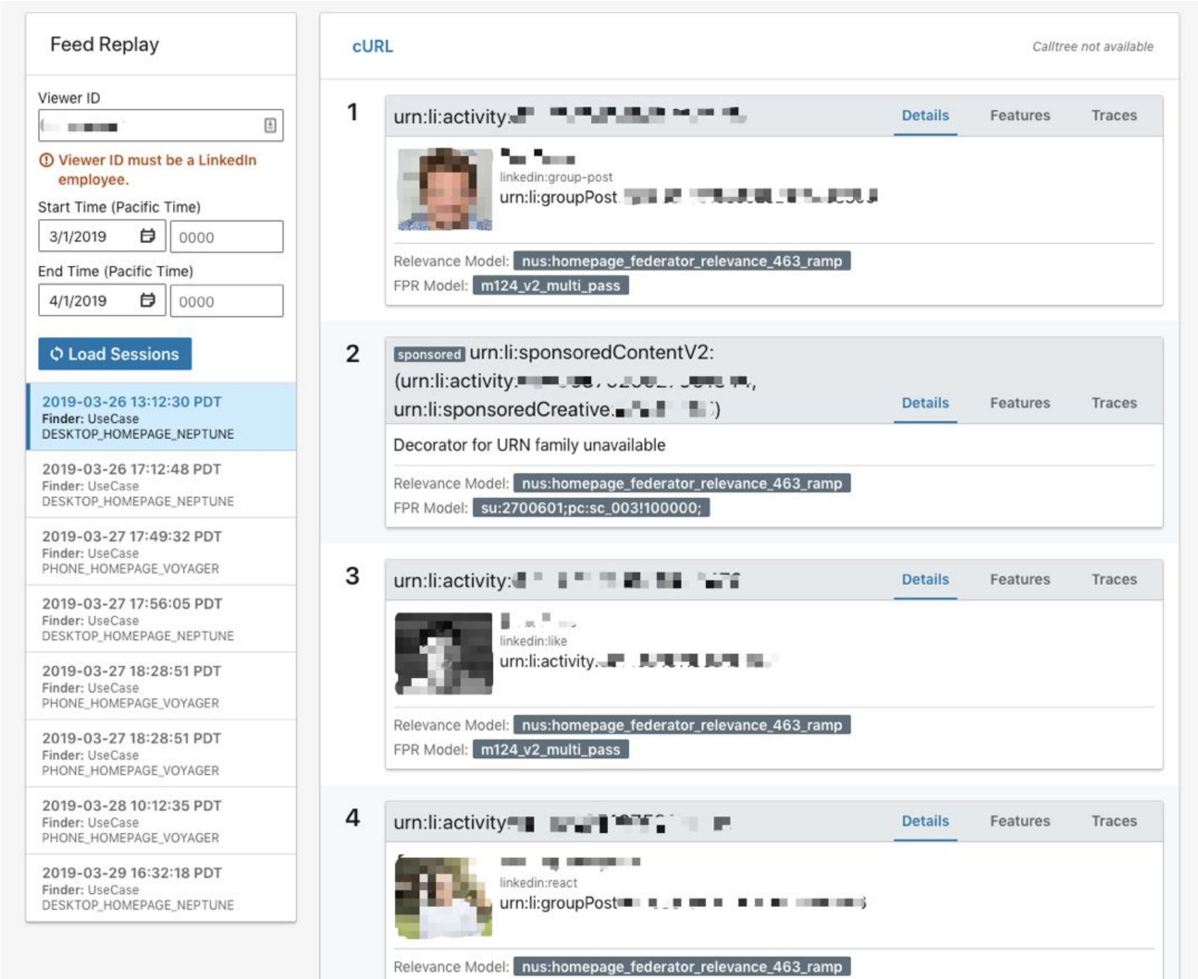
Holistic Comparison



Granular Comparison



Replay



Feed Replay
Viewer ID
Viewer ID must be a LinkedIn employee.
Start Time (Pacific Time)
3/1/2019
0000
End Time (Pacific Time)
4/1/2019
0000
Load Sessions
2019-03-26 13:12:30 PDT
Finder: UseCase
DESKTOP_HOMEPAGE_NEPTUNE
2019-03-26 17:12:48 PDT
Finder: UseCase
DESKTOP_HOMEPAGE_NEPTUNE
2019-03-27 17:49:32 PDT
Finder: UseCase
PHONE_HOMEPAGE_VOYAGER
2019-03-27 17:56:05 PDT
Finder: UseCase
DESKTOP_HOMEPAGE_NEPTUNE
2019-03-27 18:28:51 PDT
Finder: UseCase
PHONE_HOMEPAGE_VOYAGER
2019-03-27 18:28:51 PDT
Finder: UseCase
PHONE_HOMEPAGE_VOYAGER
2019-03-28 10:12:35 PDT
Finder: UseCase
PHONE_HOMEPAGE_VOYAGER
2019-03-29 16:32:18 PDT
Finder: UseCase
DESKTOP_HOMEPAGE_NEPTUNE
cURL
Calltree not available
1
urn:li:activity:
Details
Features
Traces
linkedin:group-post
urn:li:groupPost
Relevance Model: nus:homepage_federator_relevance_463_ramp
FPR Model: m124_v2_multi_pass
2
sponsored urn:li:sponsoredContentV2:
(urn:li:activity:
urn:li:sponsoredCreative:
Details
Features
Traces
Decorator for URN family unavailable
Relevance Model: nus:homepage_federator_relevance_463_ramp
FPR Model: su:2700601;pc:sc_003!100000;
3
urn:li:activity:
Details
Features
Traces
linkedin:like
urn:li:activity:
Relevance Model: nus:homepage_federator_relevance_463_ramp
FPR Model: m124_v2_multi_pass
4
urn:li:activity
Details
Features
Traces
linkedin:react
urn:li:groupPost
Relevance Model: nus:homepage_federator_relevance_463_ramp

Teams

- Search
- Feed
- Comments
- People you may know
- Jobs you may be interested in
- Notification

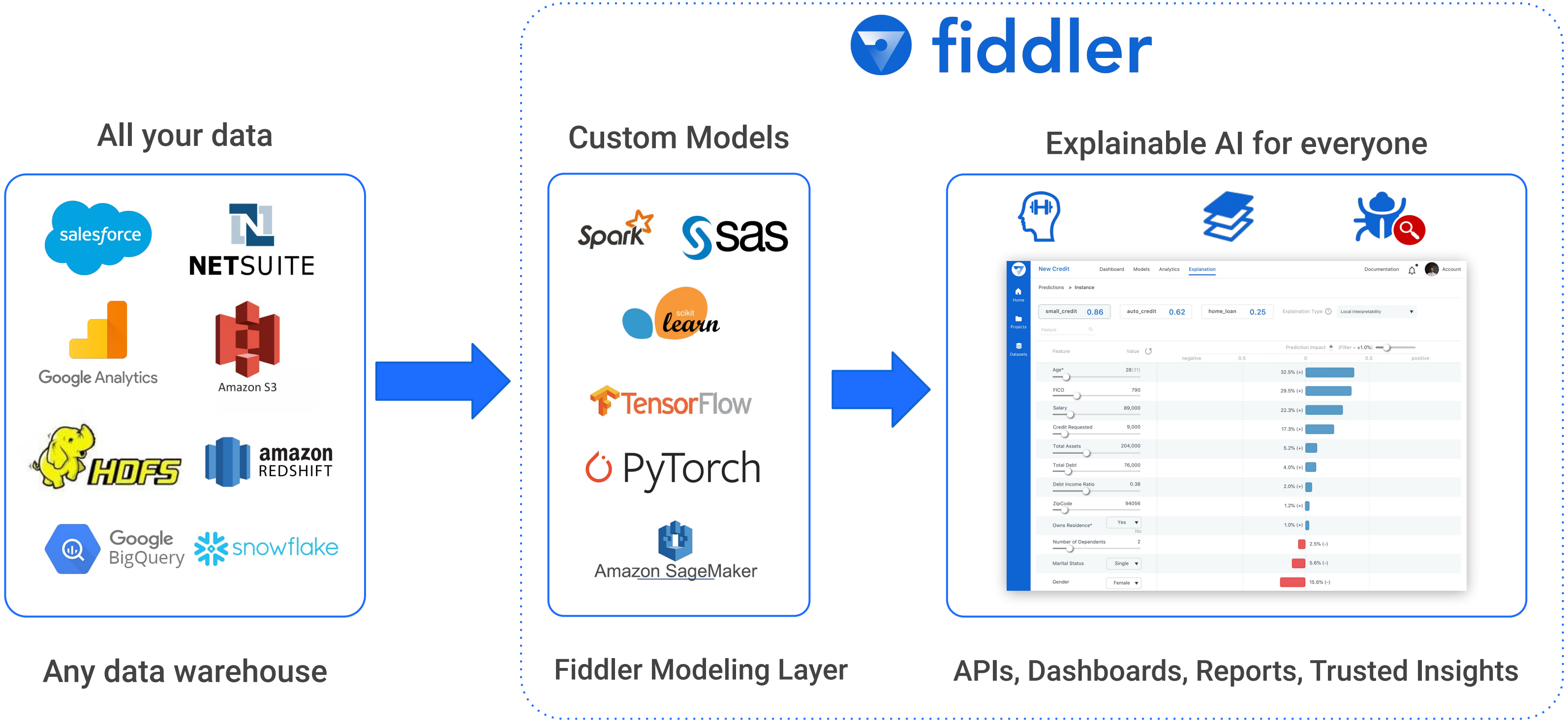
Case Study:

Building an Explainable AI Engine @ fiddler

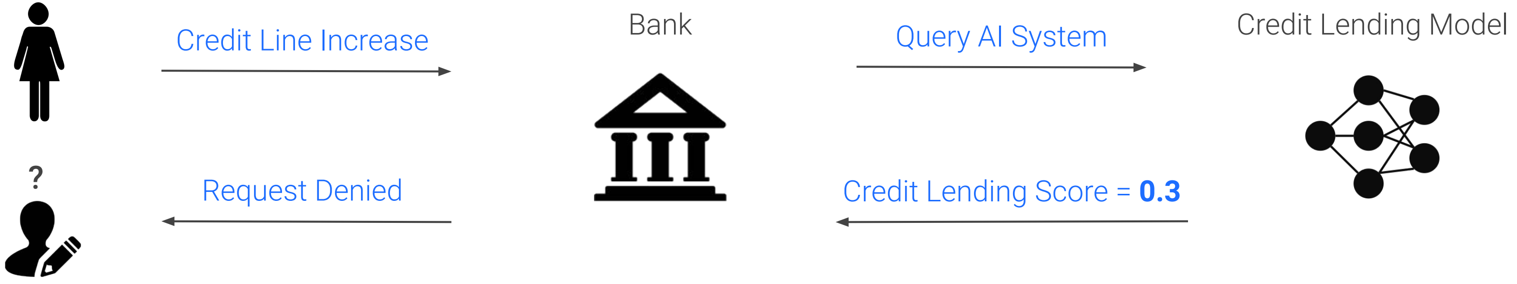
Luke Merrick

Fiddler’s Explainable AI Engine

Mission: Unlock **Trust, Visibility and Insights** by making **AI Explainable** in every enterprise



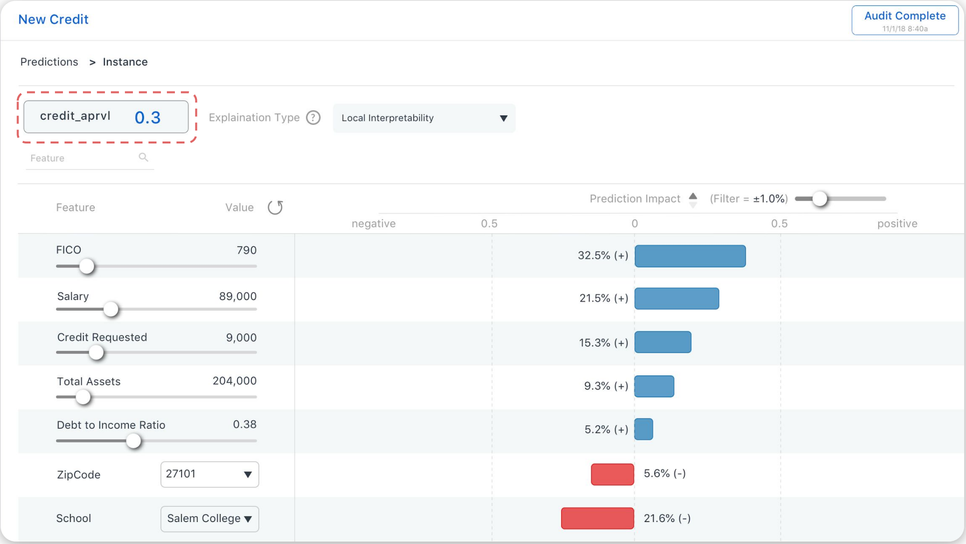
Example: Credit Lending in a black-box ML world



Why? Why not? How?

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

Explain individual predictions (using Shapley Values)



How Can This Help…

Customer Support

Why was a customer loan rejected?

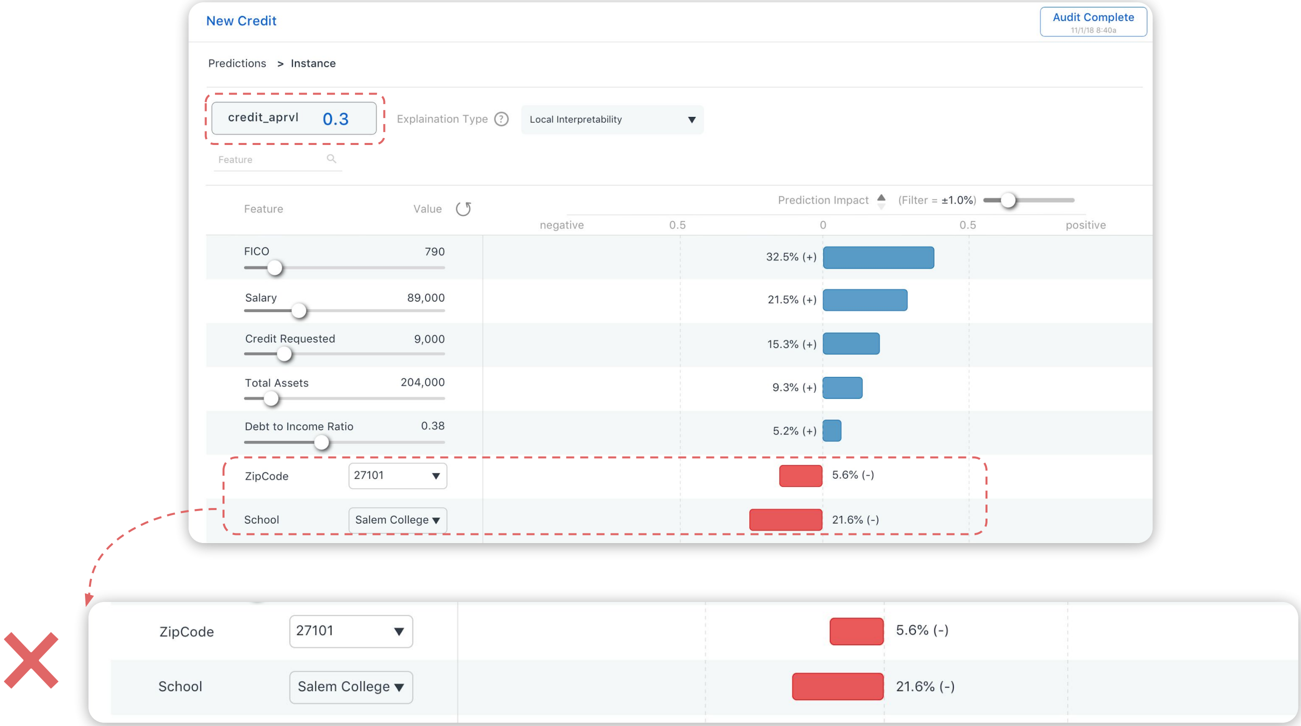
Bias & Fairness

How is my model doing across demographics?

Lending LOB

What variables should they validate with customers on "borderline" decisions?

Explain individual predictions (using Shapley Values)



How Can This Help…

Customer Support

Why was a customer loan rejected?

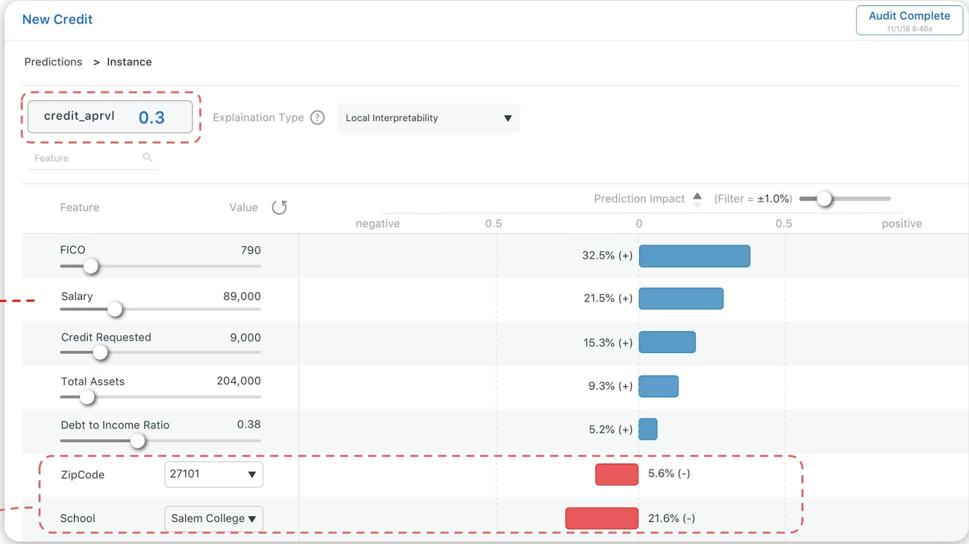
Bias & Fairness

How is my model doing across demographics?

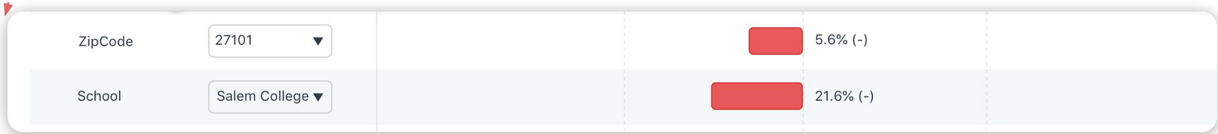
Lending LOB

What variables should they validate with customers on “borderline” decisions?

Explain individual predictions (using Shapley Values)



Probe the model on **counterfactuals**



How Can This Help…

Customer Support

Why was a customer loan rejected?

Bias & Fairness

How is my model doing across demographics?

Lending LOB

What variables should they validate with customers on “borderline” decisions?

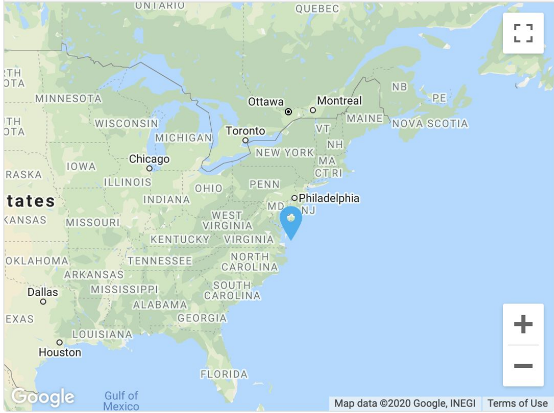
Integrating explanations

Debt Consolidation Loan

debt_consolidation

Need this loan for credit card debt consolidation!!! The fixed rate on this loan will help bring multiple payments to only one lower monthly payment.

Request Location



Repayment Model

Repayment probability: 54.4%

Fiddler Explanations

Model Feature	Value	Feature Impact
loan_amnt	8250	42%
pub_rec_bankruptcies	1	-3%
home_ownership	MORTGAGE	13%
emp_length	10+ years	3%
annual_inc	50000	-15%
revol_bal	4544	-7%
revol_util	79.7	-16%
delinq_2yrs	0	2%

Powered by fiddler

Record ID: 6

Previous Next

How Can This Help...

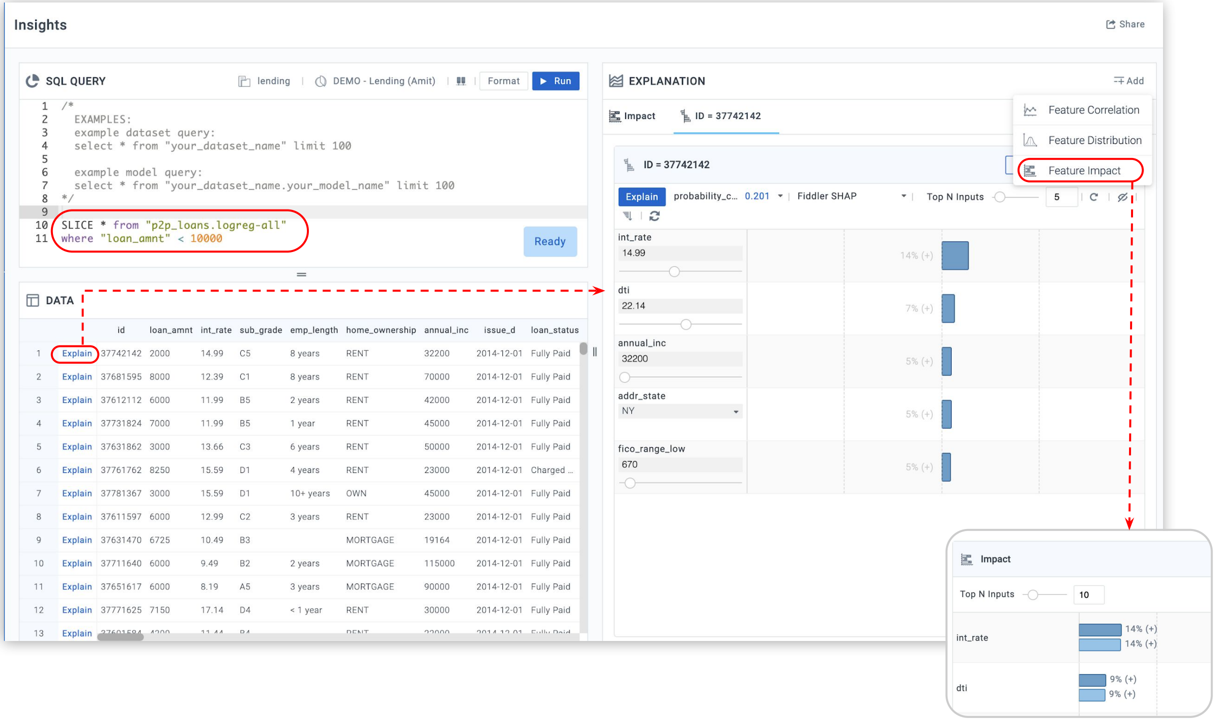
Customer Support

Why was a customer loan rejected?

Why was the credit card limit low?

Why was this transaction marked as fraud?

Slice & Explain



How Can This Help…

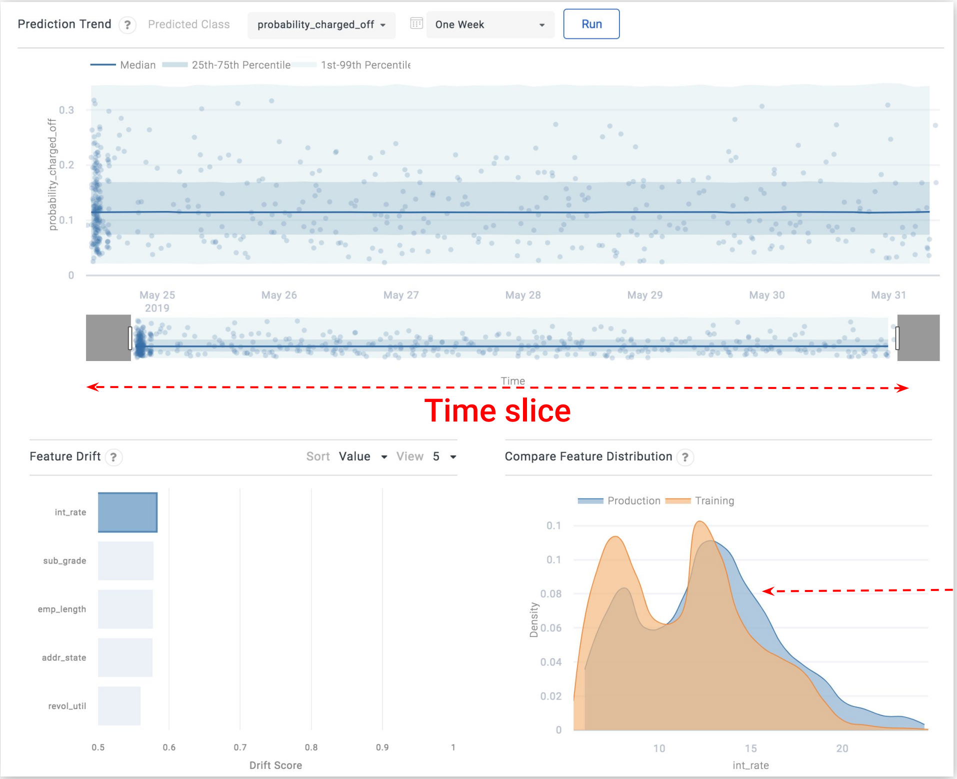
Global Explanations

What are the primary feature drivers of the dataset on my model?

Region Explanations

How does my model perform on a certain slice? Where does the model not perform well? Is my model uniformly fair across slices?

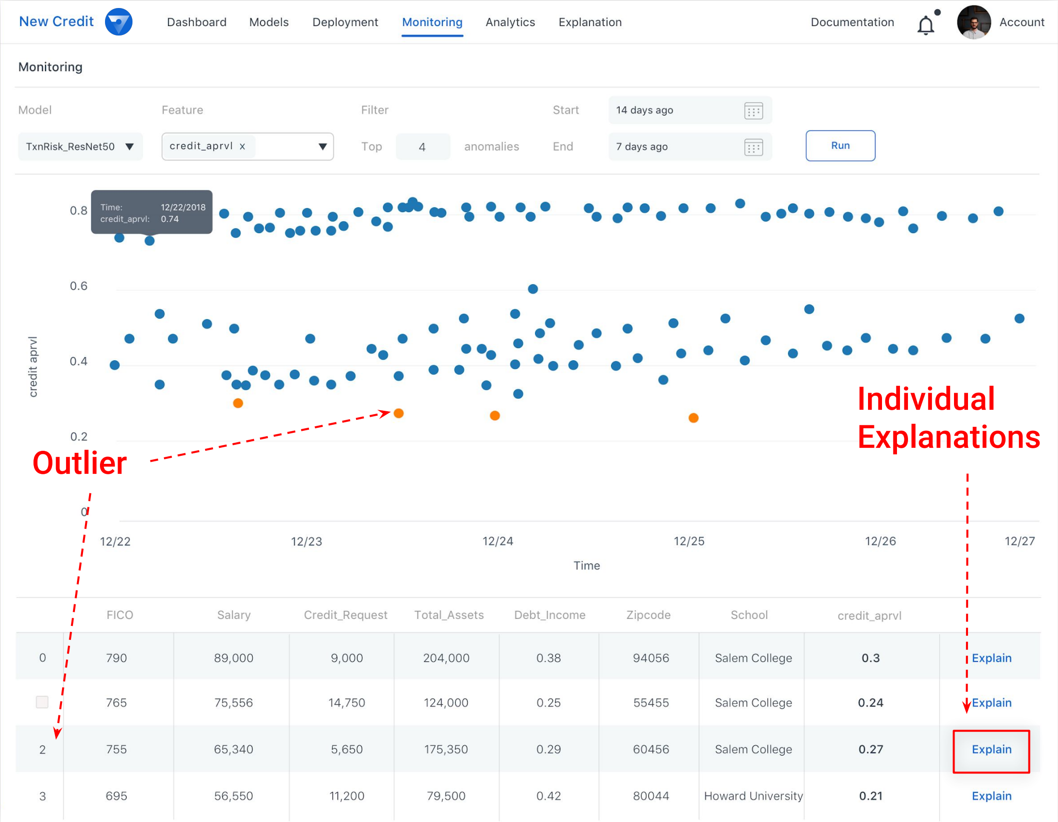
Model Monitoring: Feature Drift



Feature distribution for time slice relative to training distribution

Investigate Data Drift Impacting Model Performance

Model Monitoring: Outliers with Explanations



How Can This Help...

Operations

Why are there outliers in model predictions? What caused model performance to go awry?

Data Science

How can I improve my ML model? Where does it not do well?

Some lessons learned at Fiddler

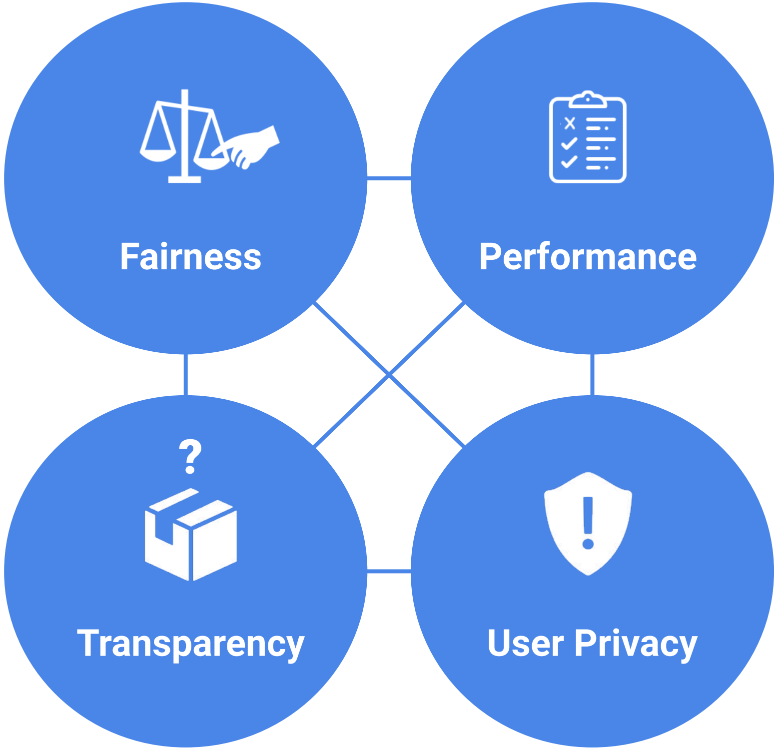
- Attributions are contrastive to their baselines
- Explaining explanations is important (e.g. good UI)
- In practice, we face engineering challenges as much as theoretical challenges

Recap

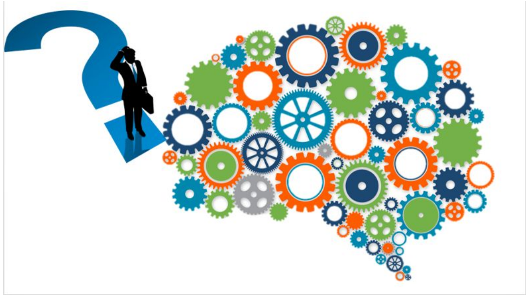
- Part I: Introduction and Motivation
  - Motivation, Definitions & Properties
  - Evaluation Protocols & Metrics
- Part II: Explanation in AI (not only Machine Learning!)
  - From Machine Learning to Knowledge Representation and Reasoning and Beyond
- Part III: Explainable Machine Learning (from a Machine Learning Perspective)
- Part IV: Explainable Machine Learning (from a Knowledge Graph Perspective)
- Part V: XAI Tools on Applications, Lessons Learnt and Research Challenges

Challenges & Tradeoffs

- Lack of standard interface for ML models makes pluggable explanations hard
- Explanation needs vary depending on the type of the user who needs it and also the problem at hand.
- The algorithm you employ for explanations might depend on the use-case, model type, data format, etc.
- There are trade-offs w.r.t. Explainability, Performance, Fairness, and Privacy.



Explainability in ML: Broad Challenges



Actionable explanations

Balance between explanations & model secrecy

Robustness of explanations to failure modes (Interaction between ML components)

Application-specific challenges
- Conversational AI systems: contextual explanations
- Gradation of explanations

Tools for explanations across AI lifecycle
- Pre & post-deployment for ML models
- Model developer vs. End user focused

Thanks! Questions?

- Feedback most welcome :-)
    - freddy.lecue@inria.fr, krishna@fiddler.ai, sgeyik@linkedin.com, kenthk@amazon.com, vamithal@linkedin.com, ankur@fiddler.ai, luke@fiddler.ai, p.minervini@ucl.ac.uk, riccardo.guidotti@unipi.it
- Tutorial website: https://xaitutorial2020.github.io
- To try Fiddler, please send an email to info@fiddler.ai
- To try Thales XAI Platform , please send an email to freddy.lecue@thalesgroup.com

Appendix

Case Study:

LinkedIn Talent Platform

"Diversity Insights and Fairness-Aware Ranking"

Sahin Cem Geyik, Krishnaram Kenthapadi

Guiding Principle: "Diversity by Design"

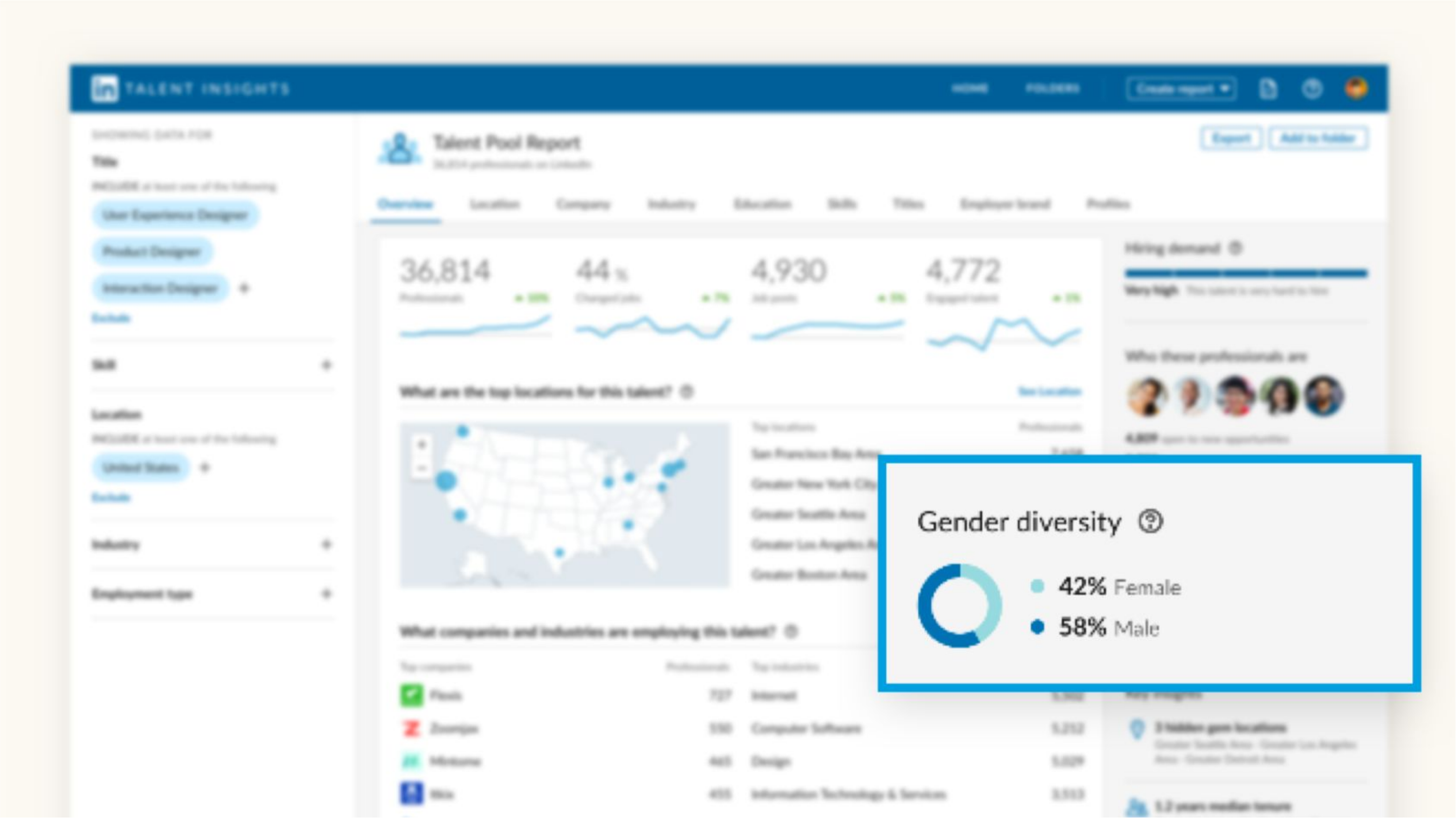
“Diversity by Design” in LinkedIn’s Talent Solutions

Insights to Identify Diverse Talent Pools

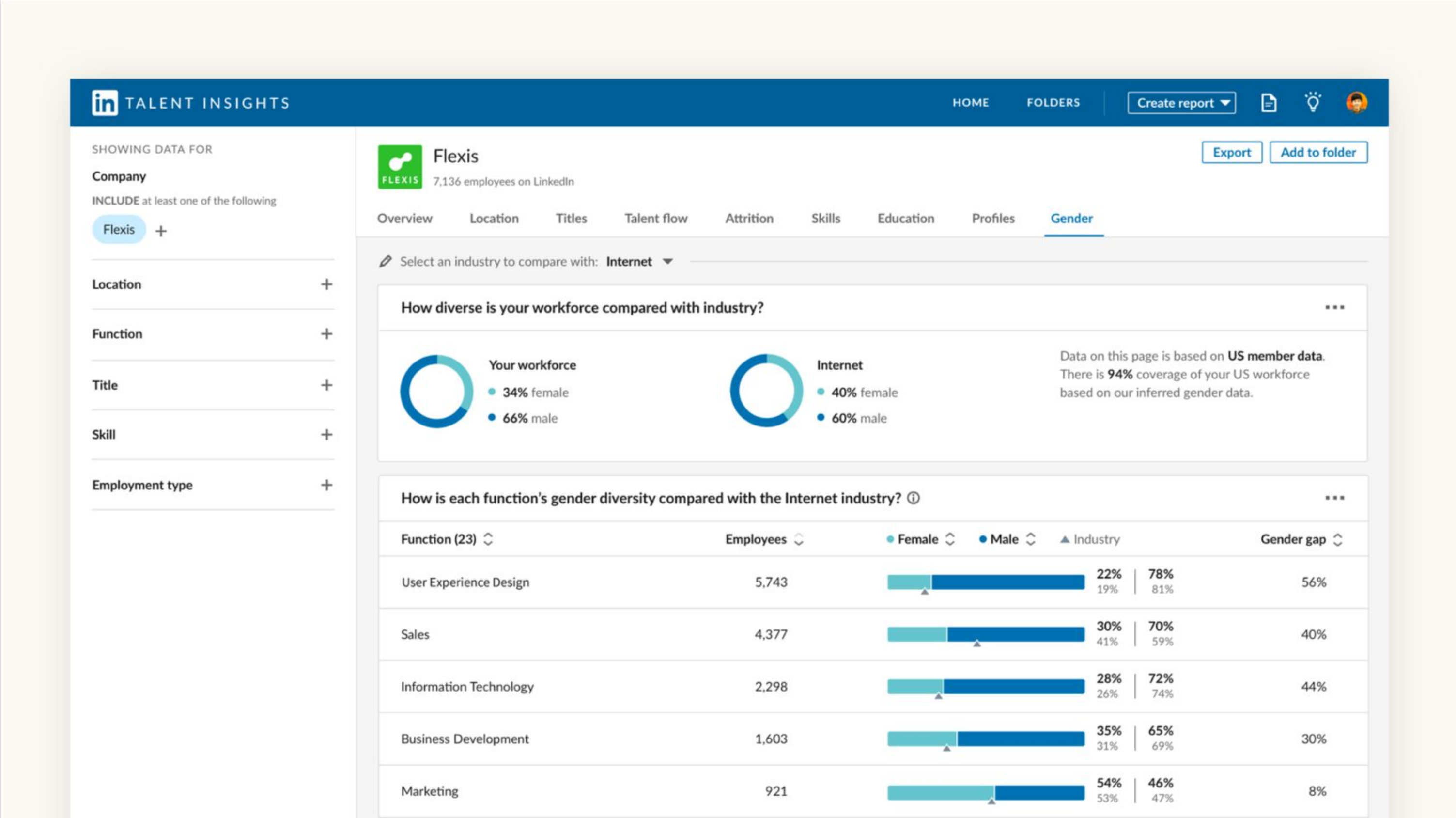
Representative Talent Search Results

Diversity Learning Curriculum

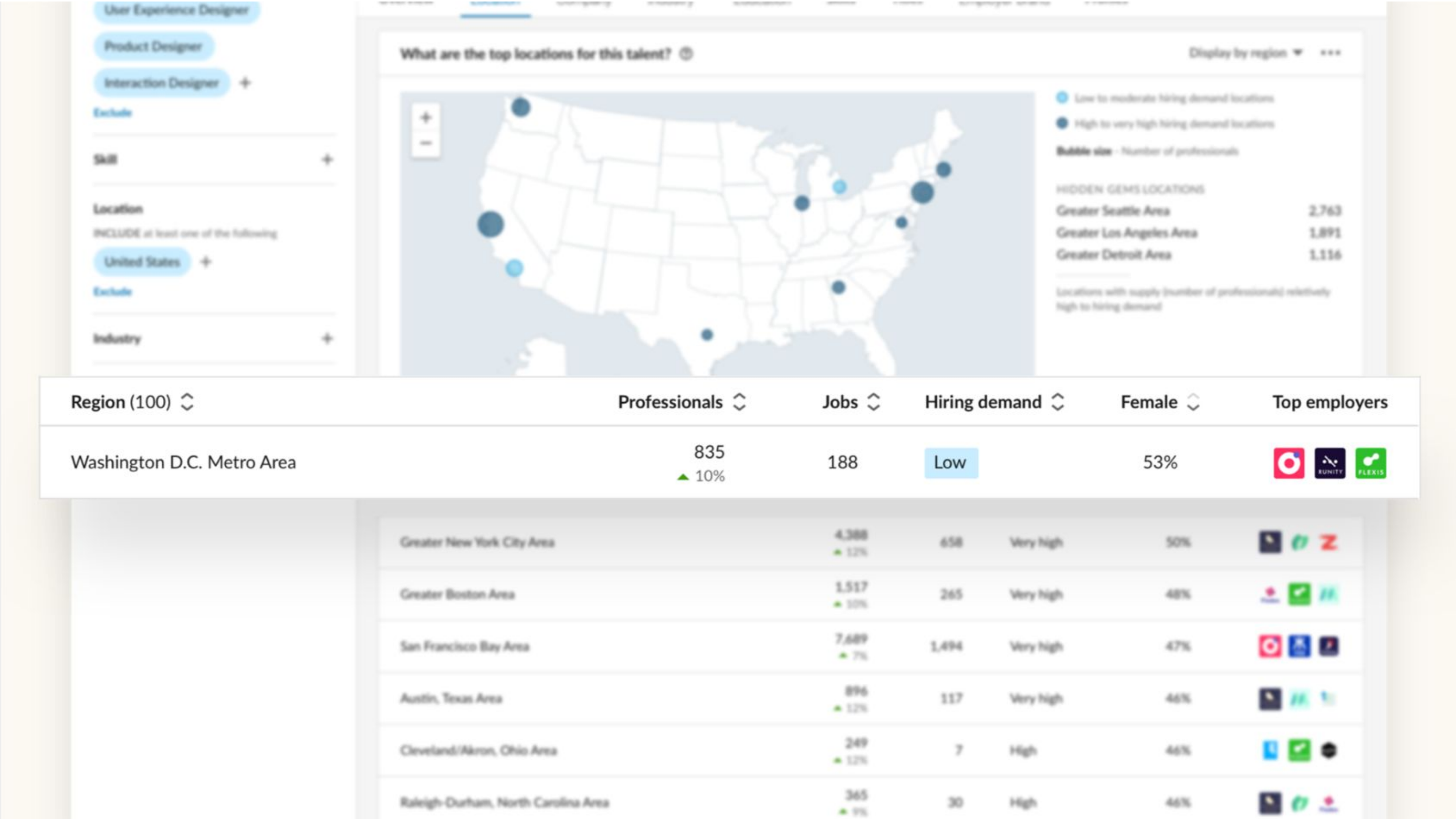
Plan for Diversity



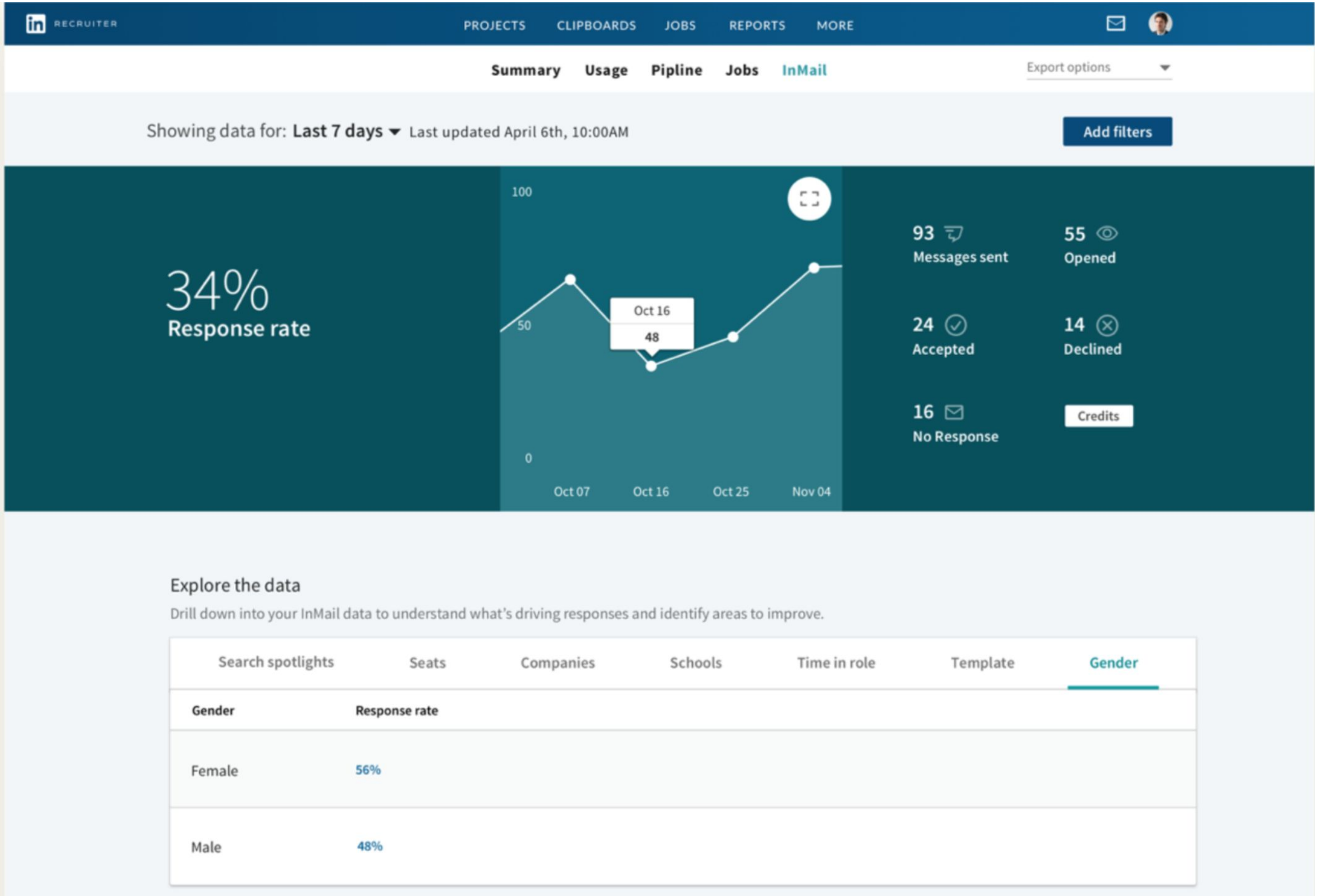
Plan for Diversity



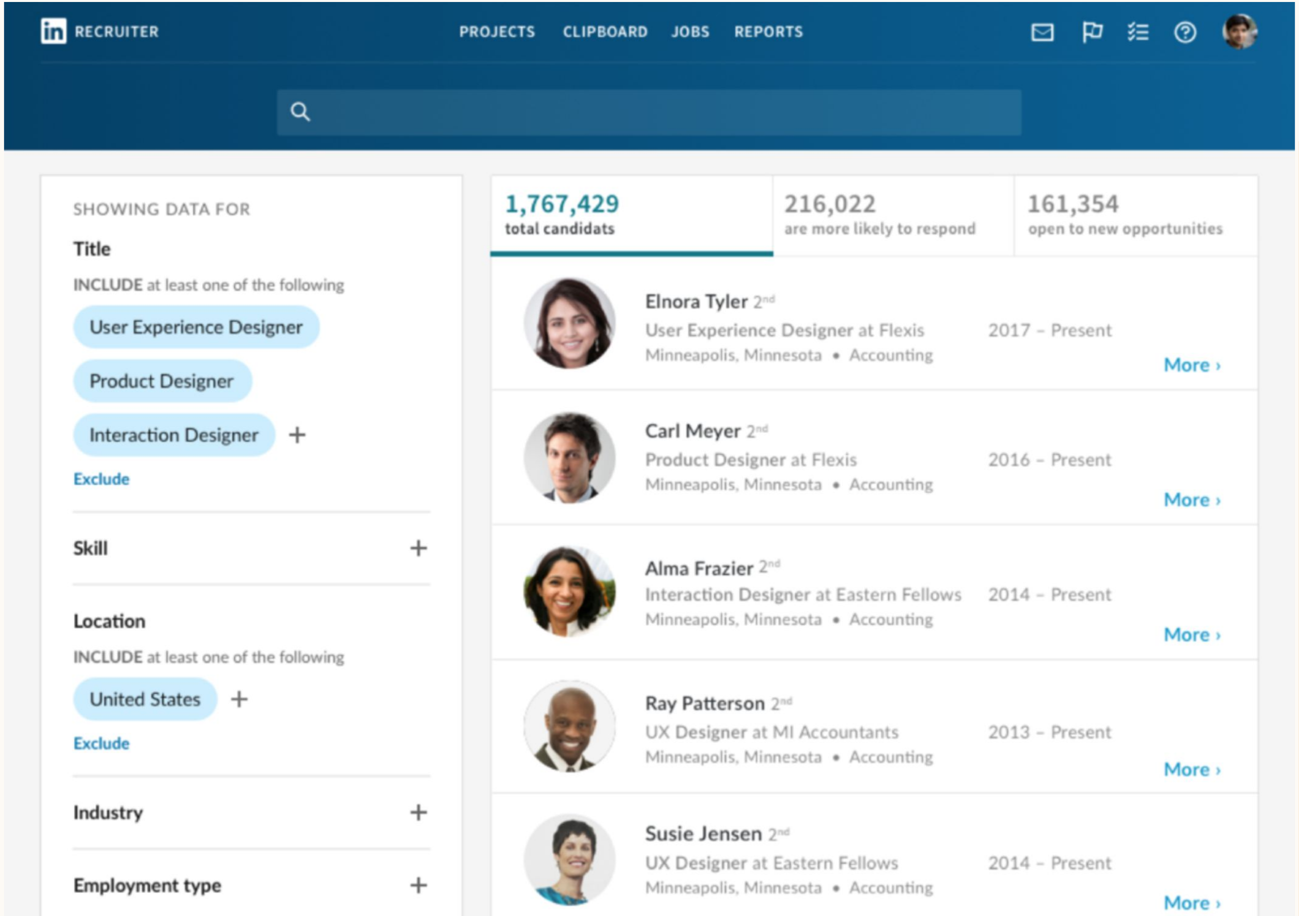
Identify Diverse Talent Pools



Inclusive Job Descriptions / Recruiter Outreach



Representative Ranking for Talent Search



S. C. Geyik, S. Ambler, K. Kenthapadi, Fairness-Aware Ranking in Search & Recommendation Systems with Application to LinkedIn Talent Search, KDD'19.

[Microsoft's AI/ML conference (MLADS'18). **Distinguished Contribution Award**]

Building Representative Talent Search at LinkedIn (LinkedIn engineering blog)

Intuition for Measuring and Achieving Representativeness

Ideal: Top ranked results should follow a desired distribution on gender/age/…

E.g., same distribution as the underlying talent pool

Inspired by "Equal Opportunity" definition [Hardt et al, NIPS'16]

Defined measures (skew, divergence) based on this intuition

Desired Proportions within the Attribute of Interest

Compute the proportions of the values of the attribute (e.g., gender, gender-age combination) amongst the set of qualified candidates

- “Qualified candidates” = Set of candidates that match the search query criteria
- Retrieved by LinkedIn’s Galene search engine

Desired proportions could also be obtained based on legal mandate / voluntary commitment

Fairness-aware Reranking Algorithm (Simplified)

Partition the set of potential candidates into different buckets for each attribute value

Rank the candidates in each bucket according to the scores assigned by the machine-learned model

Merge the ranked lists, balancing the representation requirements and the selection of highest scored candidates

- Representation requirement: Desired distribution on gender/age/…
- Algorithmic variants based on how we achieve this balance

Validating Our Approach

Gender Representativeness

- Over 95% of all searches are representative compared to the qualified population of the search

Business Metrics

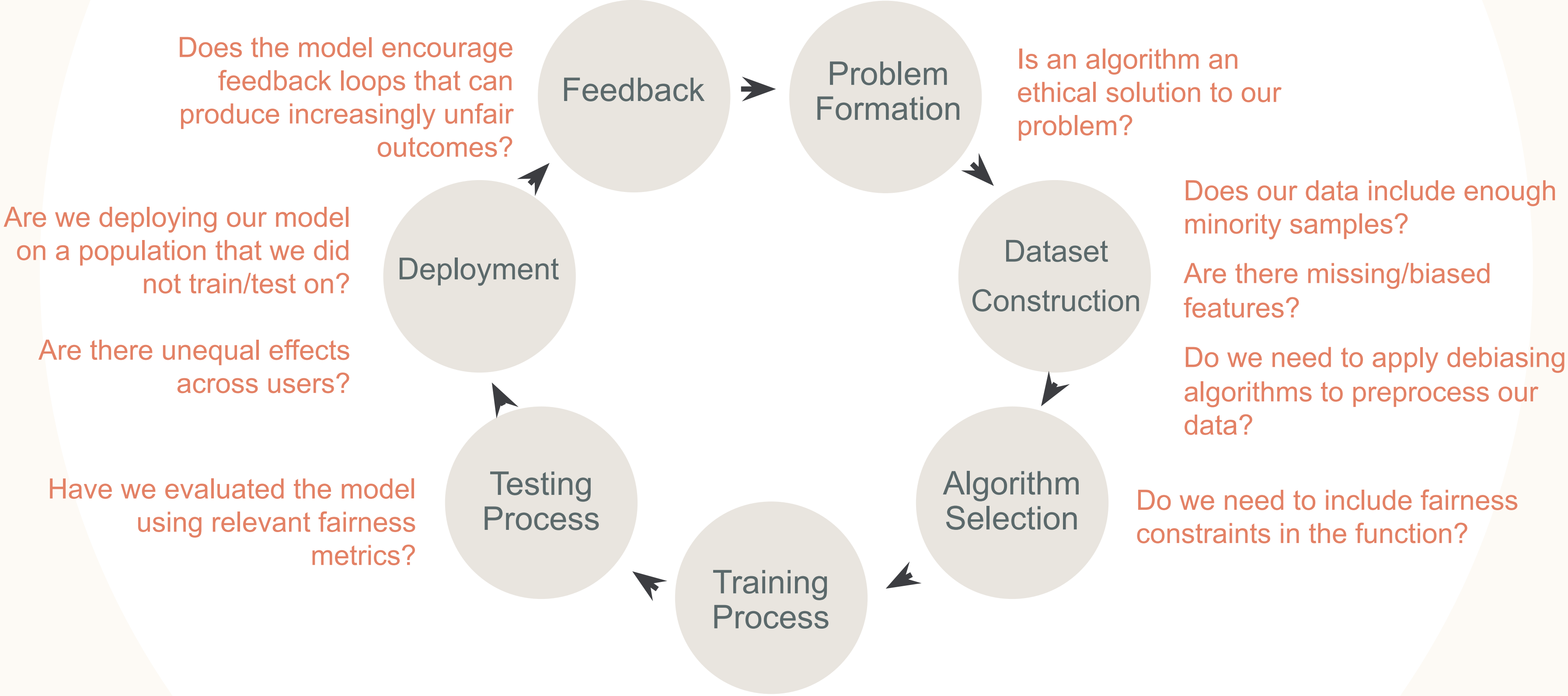
- A/B test over LinkedIn Recruiter users for two weeks
- No significant change in business metrics (e.g., # InMails sent or accepted)

Ramped to 100% of LinkedIn Recruiter users worldwide

Lessons learned

- Post-processing approach desirable
  - Model agnostic
    - Scalable across different model choices for our application
  - Acts as a “fail-safe”
    - Robust to application-specific business logic
  - Easier to incorporate as part of existing systems
    - Build a stand-alone service or component for post-processing
    - No significant modifications to the existing components
  - Complementary to efforts to reduce bias from training data & during model training
- Collaboration/consensus across key stakeholders

Engineering for Fairness in AI Lifecycle



Credit: K. Browne & J. Draper

Fairness

Privacy

Related AAAI'20 sessions:

1. Tutorial: Fairness-Aware Machine Learning: Practical Challenges and Lessons Learned (Sun)
2. Workshop: Explainable AI/ML (XAI) for Accountability, Fairness, and Transparency (Mon)
3. Social Impact Workshop (Wed, 8:15 – 11:45)
4. Keynote: Cynthia Rudin, Do Simpler Models Exist and How Can We Find Them? (Thu, 8 - 9am)
5. Several papers on fairness (e.g., ADS7 (Thu, 10-12), ADS9 (Thu, 1:30-3:30))
6. Research Track Session RT17: Interpretability (Thu, 10am - 12pm)

Transparency

Explainability

Algorithmic Bias

- Ethical challenges posed by AI systems

- Inherent biases present in society
  - Reflected in training data
  - AI/ML models prone to amplifying such biases
    - ACM FAT* conference / KDD’16 & NeurIPS’17 Tutorials

Image results: "Unprofessional hair for work"

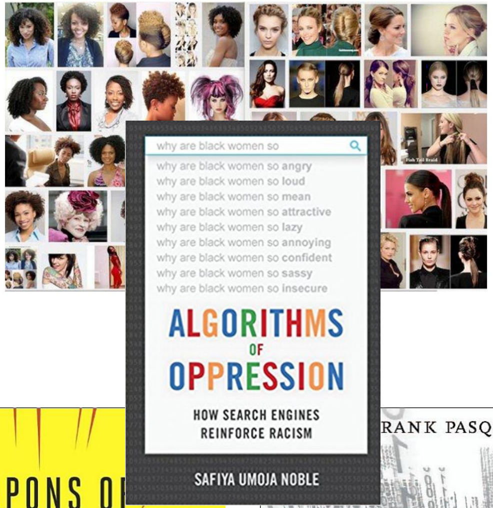
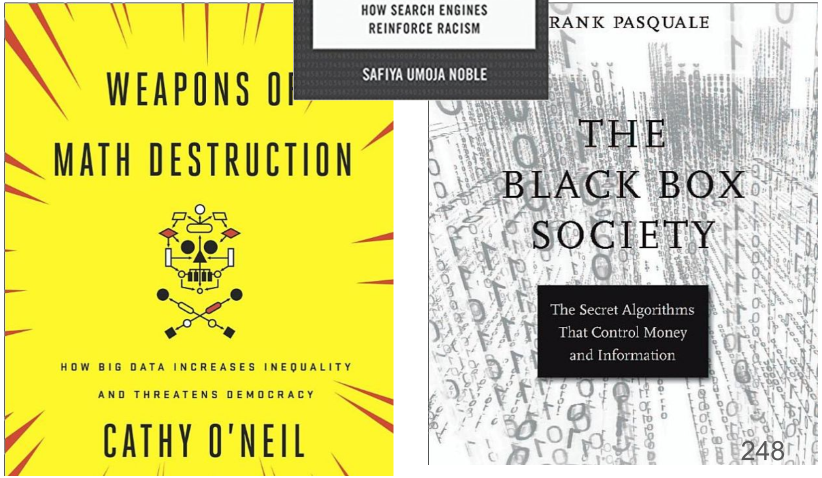
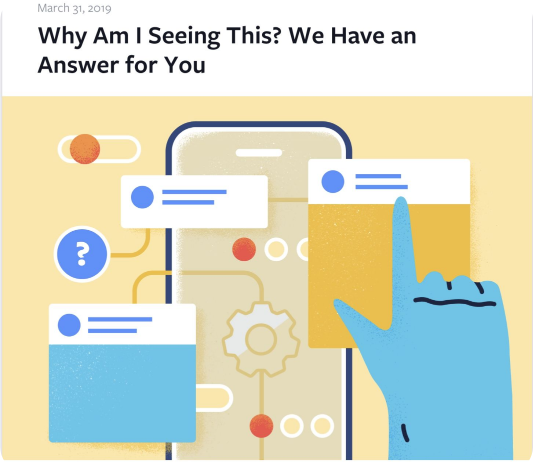
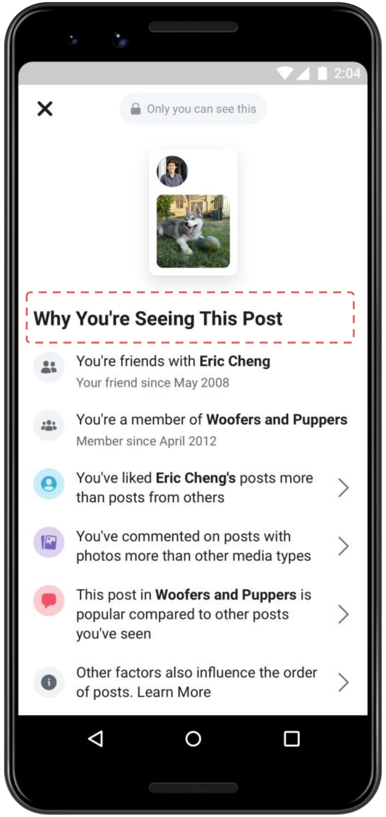


Image results: "Professional hair for work"



Example: Facebook adds Explainable AI to build Trust





Axioms

- **Insensitivity:** A variable that has no effect on the output gets no attribution
- **Sensitivity**: If baseline and input differ in a single variable, and have different outputs, then that variable should receive some attribution
- **Linearity preservation**: Attributions(α*F1 + ß*F2) = α*Attributions(F1) + ß*Attributions(F2)
- **Implementation invariance**: Two networks that compute identical functions <u>for all inputs</u> get identical attributions
- **Completeness**: Sum(attributions) = F(input) - F(baseline)
- **Symmetry**: Symmetric variables with identical values get equal attributions