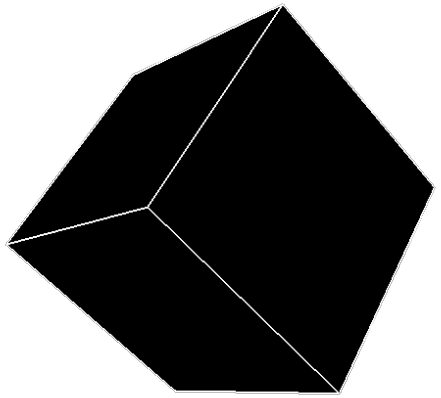
On Explainable AI:

From Theory to Motivation, Applications and Limitations

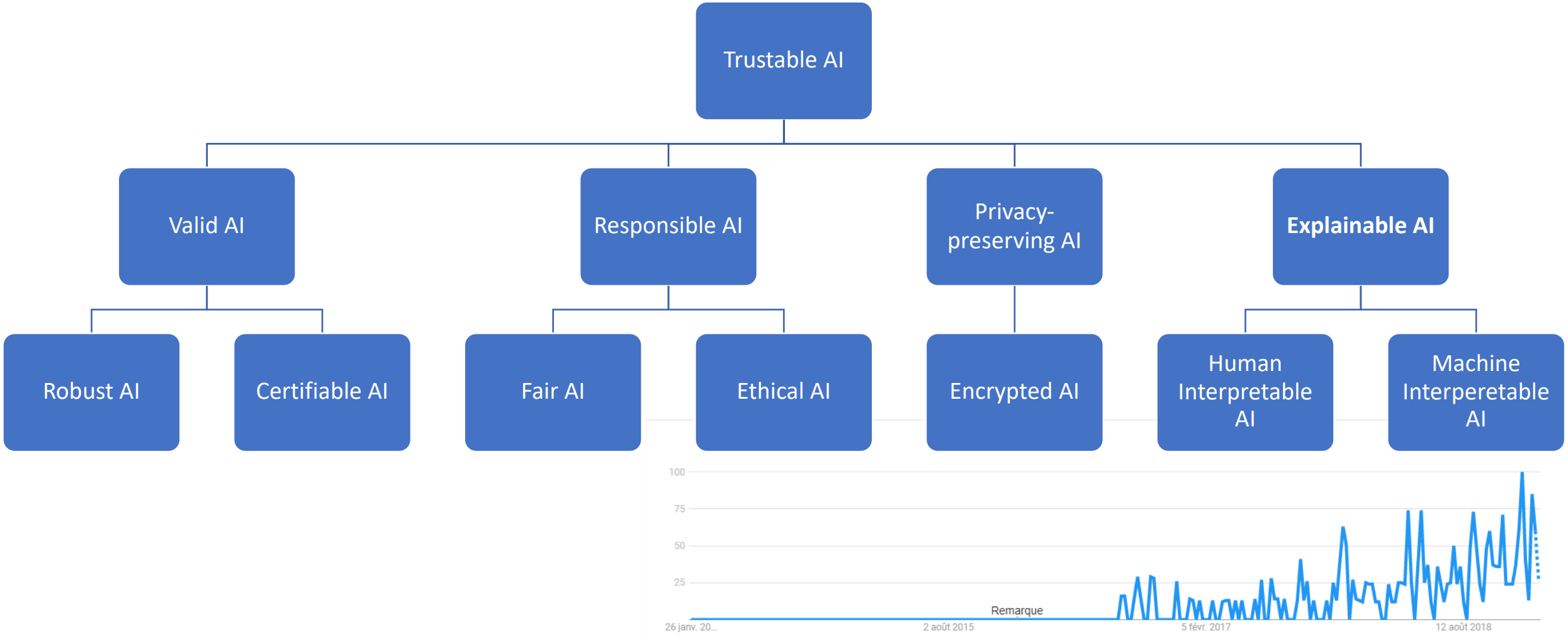
Freddy Lécué
Inria, France
CortAIx@Thales, Canada
@freddylecue

Pasquale Minervini
University College London
@PMinervini



Scope

*AI Context for Industrial Adoption



Disclaimer

- **As MANY interpretations as research areas** (check out work in Machine Learning vs Reasoning community)

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

Explanation in AI

Explanation in AI aims to create a suite of techniques that produce more explainable models, while maintaining a high level of searching, learning, planning, reasoning performance: optimization, accuracy, precision; and enable human users to understand, appropriately trust, and effectively manage the emerging generation of AI systems .

Outline

Tutorial Outline (1)

- **Explanation in Artificial Intelligence** **9:00 – 10:00**
  - Motivation
  - Definitions & Properties
  - Explanations in Different AI fields
  - The Role of Humans
  - Evaluation Protocols & Metrics
- **Explanation in Machine Learning** **10:00 – 11:00**
  - Explanation Taxonomy
  - Explanation in Machine Learning
- *Break* **11:00 – 11:30**

Tutorial Outline (2)

- **On the Role of Knowledge Graph in Explainable AI** **11:30 – 12:30**
  - Knowledge Graphs
  - Extending Machine Learning Systems with Knowledge Graphs
- *Break* **12:30 – 13:30**
- **On the Role of Reasoning in Explainable AI** **13:30 – 15:30**
  - Relational Learning
  - On Combining Neural Networks with Logic Programming
- *Break* **15:30 – 16:00**
- **Industrial Applications of XAI** **16:00 – 17:00**
- **Conclusion + Q&A** **17:00 – 18:00**

Motivation

Business to Customer

Gary Chavez added a photo you might be in. ...

about a minute ago ·

Critical Systems


BNSF

Markets We Serve (Critical Systems)

Aerospace

Space

Ground Transportation

Defence

Security

Trusted Partner For A Safer World

But not Only Critical Systems

COMPAS recidivism black bias

Opinion

OP-ED CONTRIBUTOR

When a Computer Program Keeps You in Jail

By Rebecca Wexler

June 13, 2017

Motivation (2)

Finance:

- Credit scoring, loan approval
- Insurance quotes

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

The Big Read Artificial intelligence + Add to myFT

Insurance: Robots learn the business of covering risk

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

Save

Oliver Ralph MAY 16, 2017 24

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

Motivation (3)

Healthcare

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

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

Stanford MEDICINE | News Center

Email Tweet

Researchers say use of artificial intelligence in medicine raises ethical questions

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

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

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

Rich Caruana
Microsoft Research
rcaruana@microsoft.com

Yin Lou
LinkedIn Corporation
ylou@linkedin.com

Johannes Gehrke
Microsoft
johannes@microsoft.com

Paul Koch
Microsoft Research
paulkoch@microsoft.com

Marc Sturm
NewYork-Presbyterian Hospital
mas9161@nyp.org

Noémie Elhadad
Columbia University
noemie.elhadad@columbia.edu

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

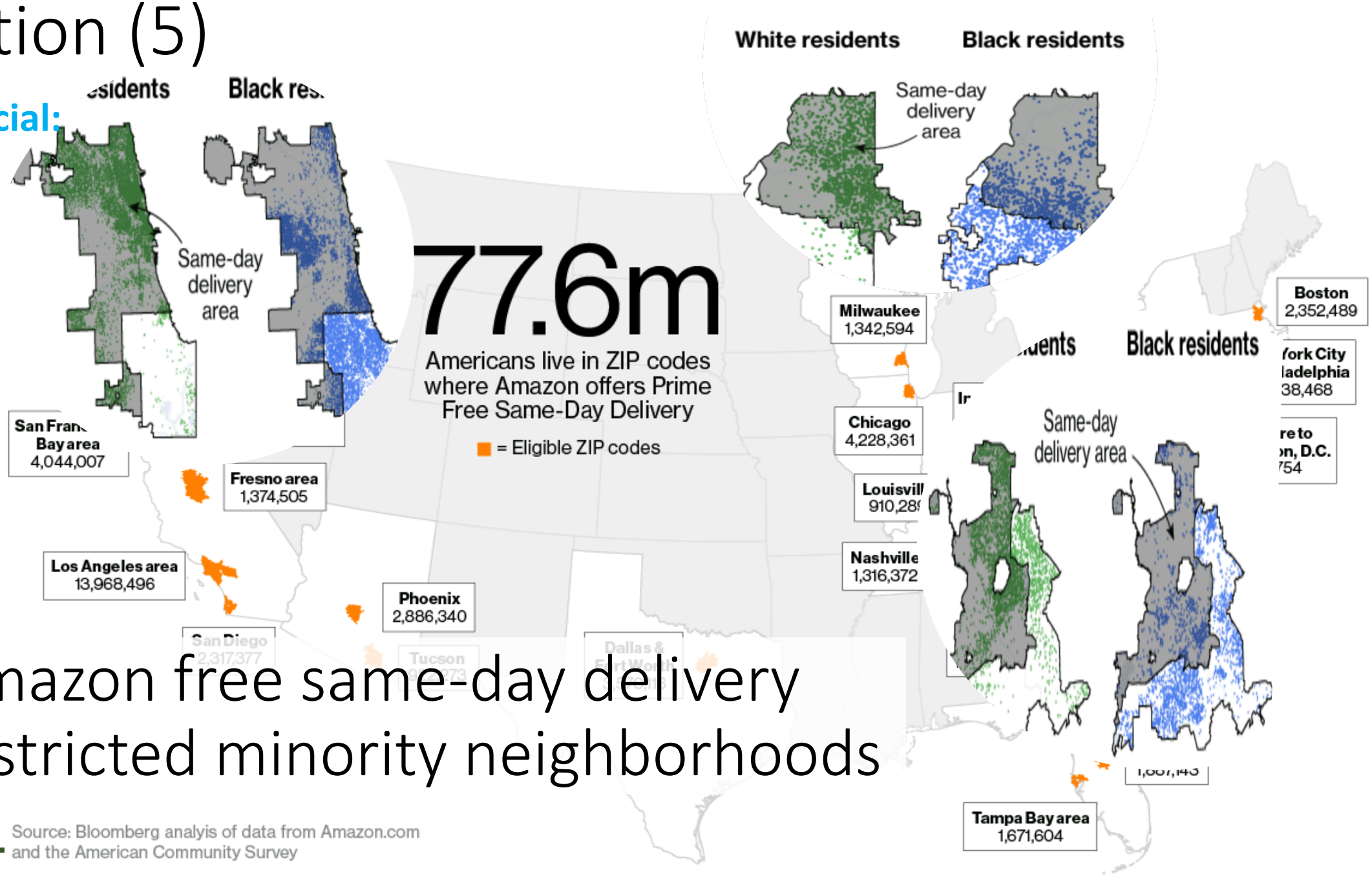
Motivation (4)

Human Resources – Talent Acquisition

- Discriminative Job Screening Software

Motivation (5)

Commercial:



No Amazon free same-day delivery for restricted minority neighborhoods

Source: Bloomberg analyis of data from Amazon.com and the American Community Survey

Industry Push for Explanation

Call for Explanation (1)

- User Acceptance & Trust

Lipton, Zachary C. "The mythos of model interpretability. Int. Conf." Machine Learning: Workshop on Human Interpretability in Machine Learning. 2016.

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

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

- Legal
  - Conformance to ethical standards, fairness
  - *Right to be informed*
  - Contestable decisions

Wachter, Sandra, Brent Mittelstadt, and Luciano Floridi. "Why a right to explanation of automated decision-making does not exist in the general data protection regulation." International Data Privacy Law 7.2 (2017): 76-99.

Goodman, Bryce, and Seth Flaxman. "European Union regulations on algorithmic decision-making and a" right to explanation"." arXiv preprint arXiv:1606.08813 (2016).

- Explanatory Debugging
  - Flawed performance metrics
  - Inadequate features
  - Distributional drift

Kulesza, Todd, et al. "Principles of explanatory debugging to personalize interactive machine learning." Proceedings of the 20th international conference on intelligent user interfaces. ACM, 2015.

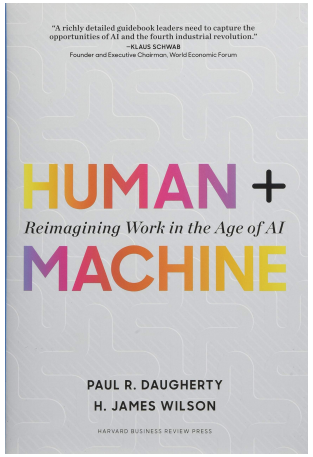
Weld, D., and Gagan Bansal. "The challenge of crafting intelligible intelligence." Communications of ACM (2018).

- Increase Insightfulness Informativeness
  - Uncovering causality

Lipton, Zachary C. "The mythos of model interpretability. Int. Conf." Machine Learning: Workshop on Human Interpretability in Machine Learning. 2016.

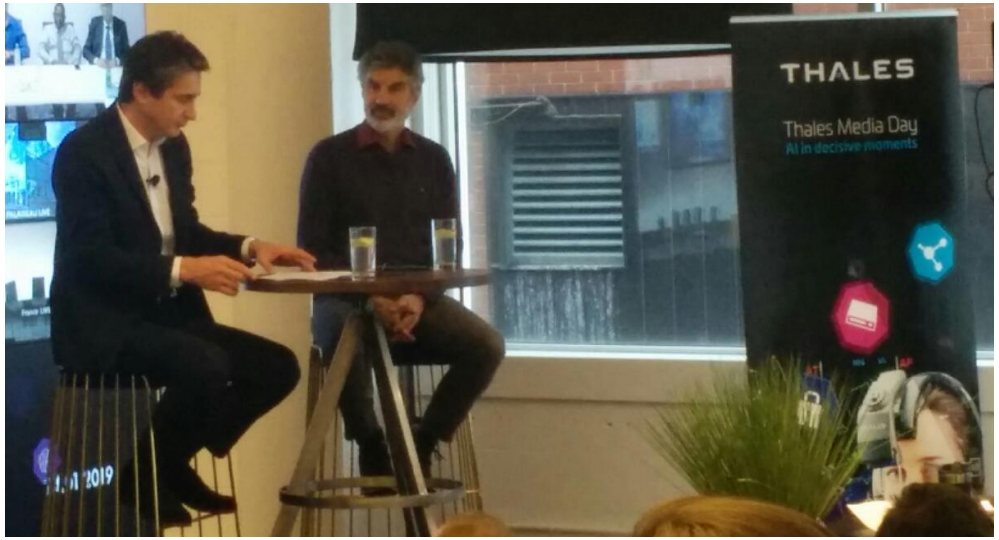
Judea Pearl: Causal Inference. NIPS Causality: Objectives and Assessment 2010: 39-58

Call for Explanation (2)



- **Critical systems / Decisive moments**
- Human factor:
  - Human decision-making affected by **greed, prejudice, fatigue, poor scalability**.
  - **Bias**
- Algorithmic decision-making on the rise.
  - More objective than humans?
  - Potentially discriminative
  - Opaque
  - Information and power asymmetry
- High-stakes scenarios = **ethical** problems!

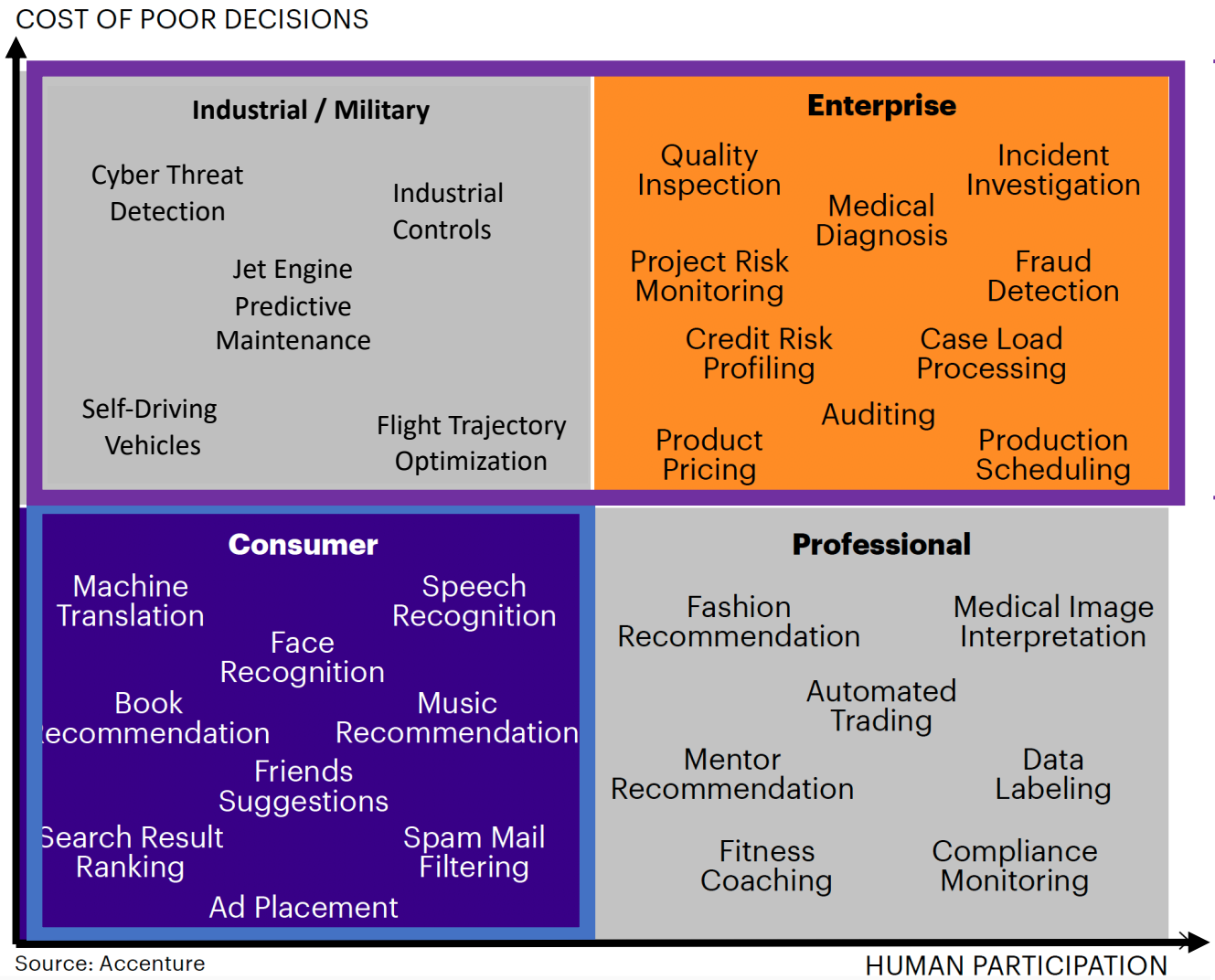
[Lepri et al. 2018]



Where is the Impact?

Trustable AI and eXplainable AI: a Reality Need

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



Most impactful successes of AI to come

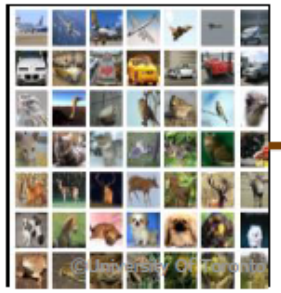
Most prominent successes of AI to date

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

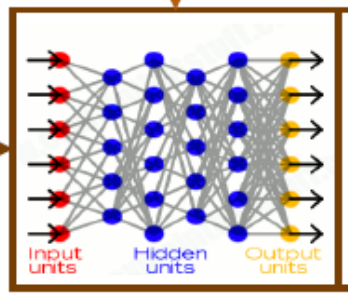
XAI in a Nutshell

XAI in a Nutshell

Today



Learning Process

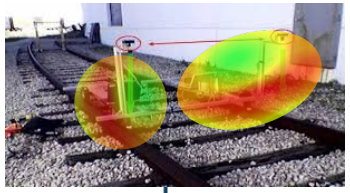


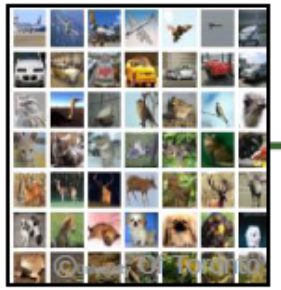
This is an obstacle on rail train

- Why did you do that?
- Why not something else?
- When do you succeed?
- When do you fail?
- When can I trust you?
- How do I correct an error?

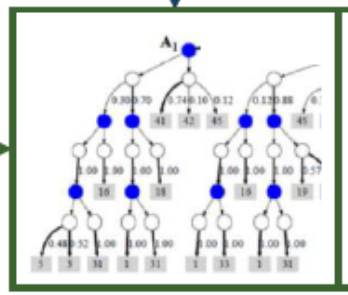
Training Data | Learned Function | Output | User with a Task

Tomorrow





New Learning Process



Obstacle on rail train

- **Obstruction covering full width**

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

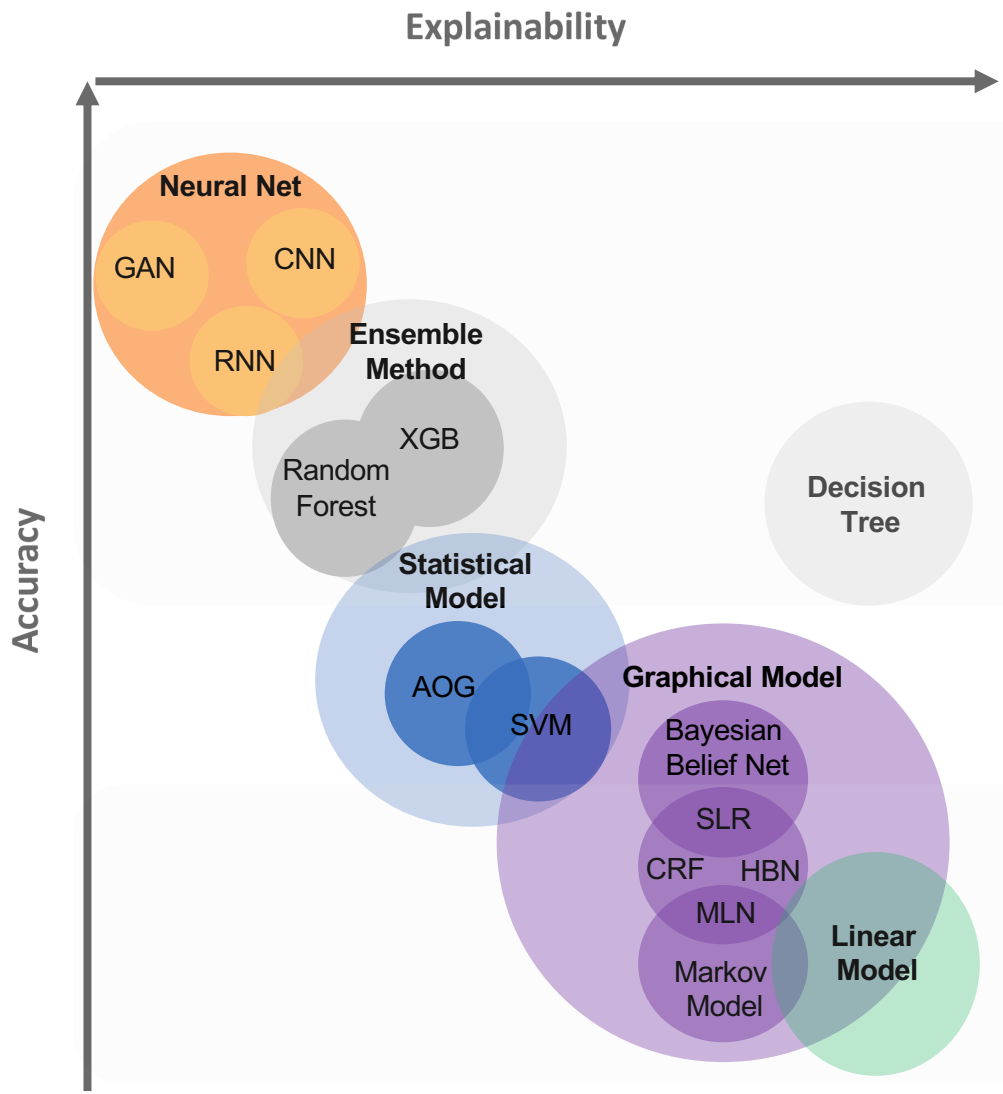
Training Data | Explainable Model | Explanation Interface | User with a Task

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

How to Explain? Accuracy vs. Explanability

Learning

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



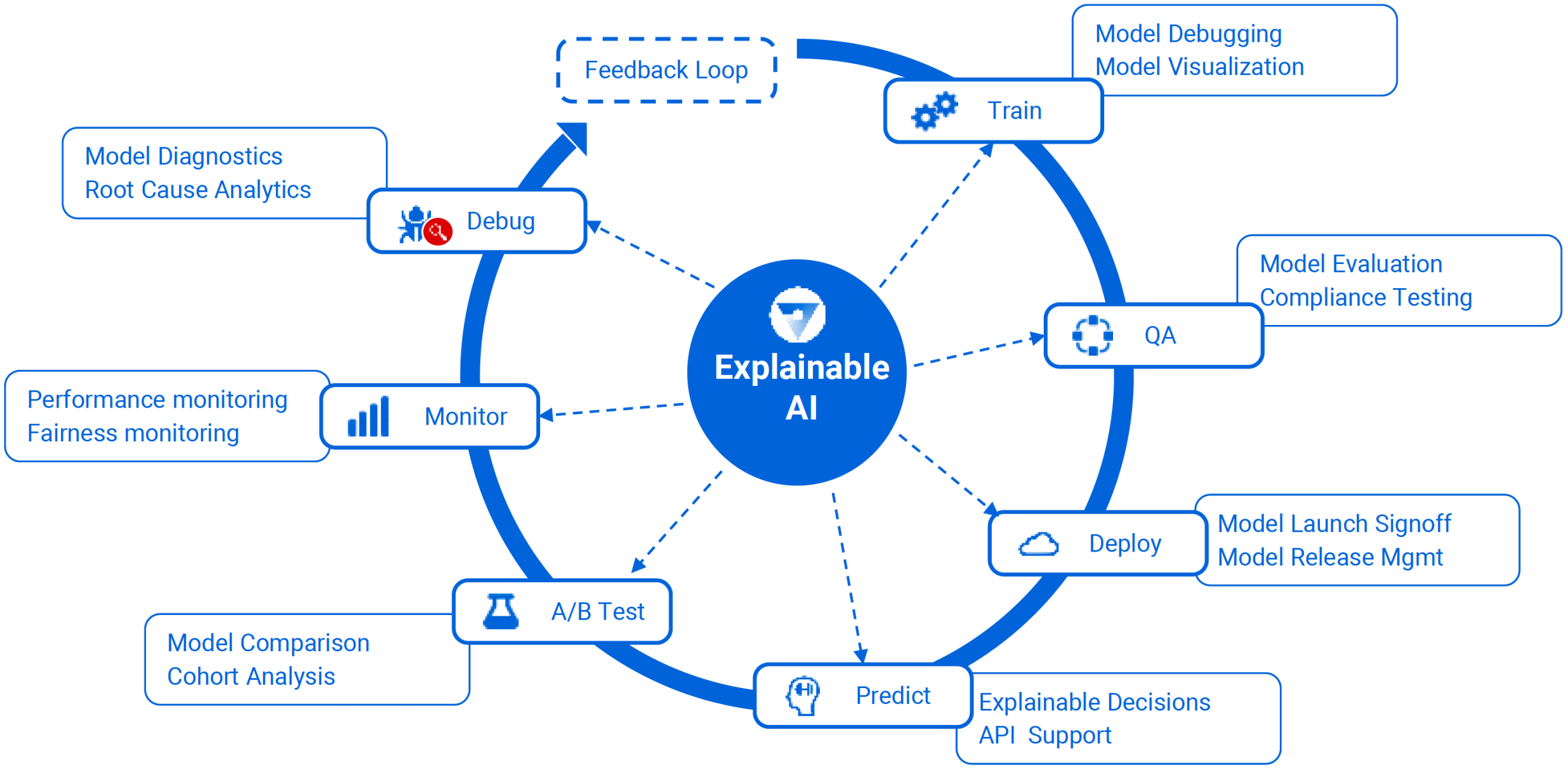
Interpretability

- Non-Linear functions
- Polynomial functions
- Quasi-Linear functions

XAI Objective

Supporting Industrialization of AI at Scale

Explainability by Design for AI Products



XAI Definitions

Explanation vs. Interpretability

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*.

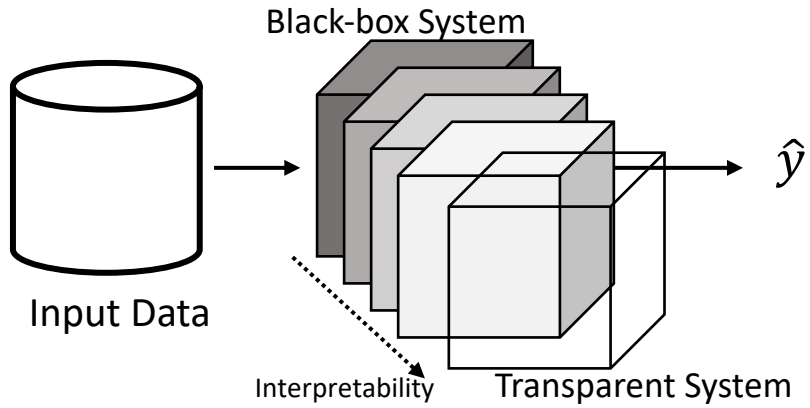
interpret | ɪn'tə:prɪt |

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

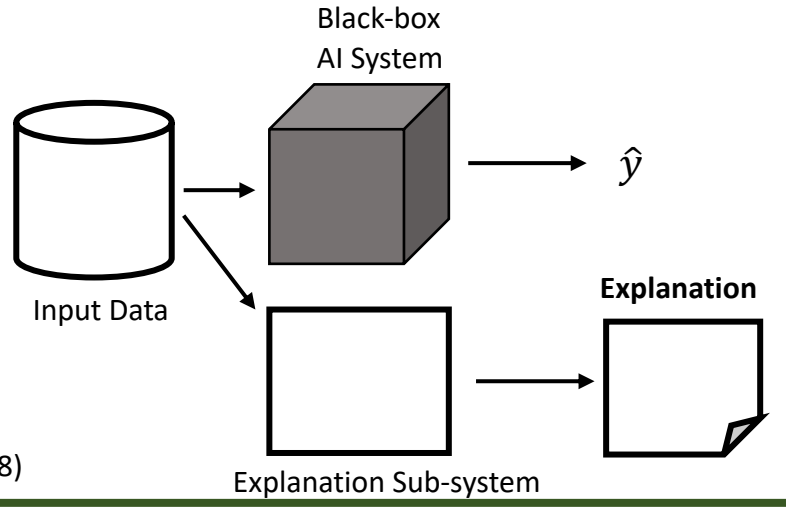
1 explain the meaning of (information or actions): *the evidence is difficult to interpret*.

Transparent Design vs Post-hoc Explanation

Transparent design reveals ***how*** **a model functions**.



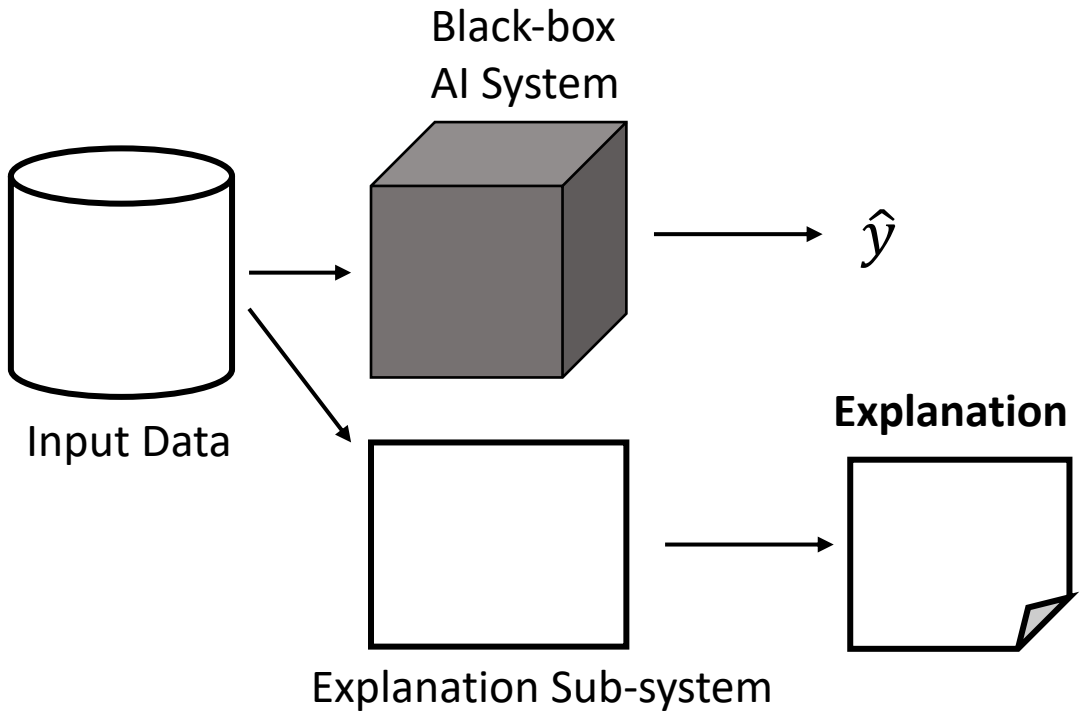
Post-hoc Explanation explains ***why*** **a black-box model behaved that way**.



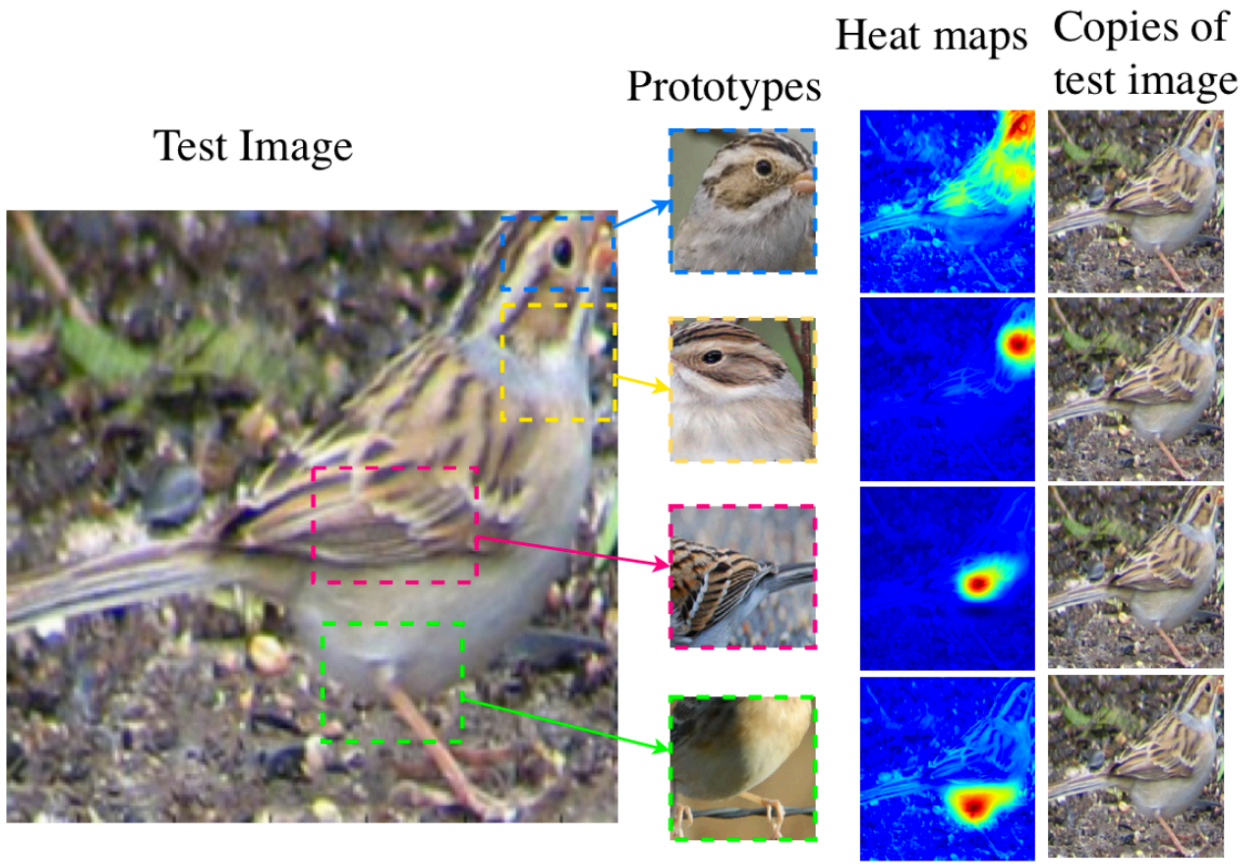
Brent D. Mittelstadt, Chris Russell, Sandra Wachter: Explaining Explanations in AI. CoRR abs/1811.01439 (2018)

So, What is an Explanation?

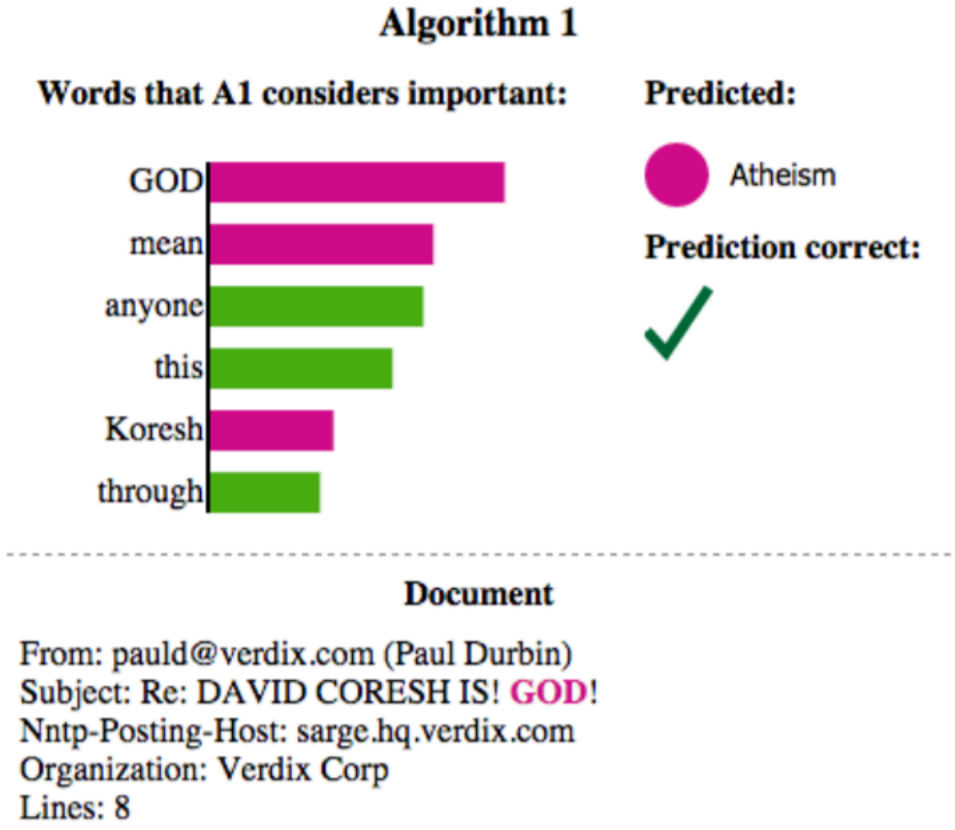
- **No formal, technical, agreed upon definition!**
- Comprehensive philosophical overview out of scope of the tutorial [Miller 2017]
- Not limited to machine learning! [Lipton 2016, Tomsett et al. 2018, Rudin 2018]



Examples



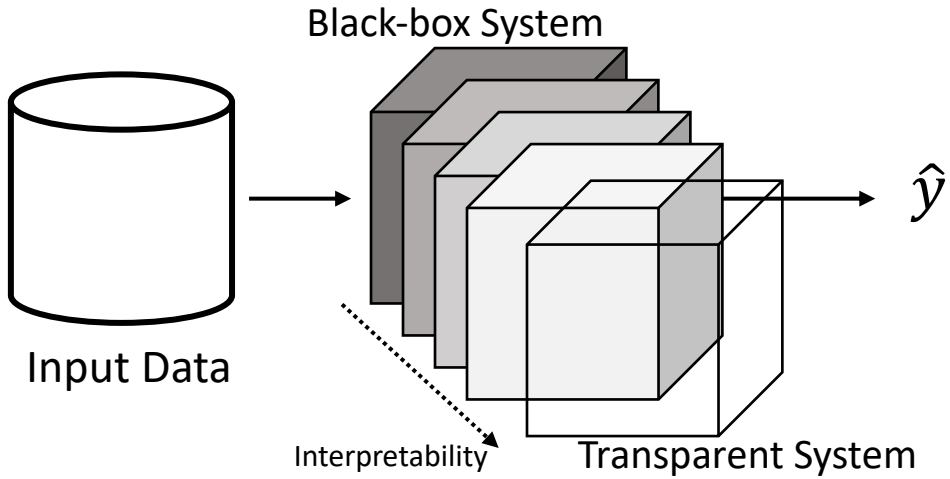
[Chen and Rudin 2018]



[Ribeiro et al. 2016]

What About Interpretability?

- Interpretability as Multi-Faceted Concept
  - Interpretability is an ill-defined term!
  - **Not** a monolithic concept



Lipton, Zachary C. "The mythos of model interpretability. Int. Conf." Machine Learning: Workshop on Human Interpretability in Machine Learning. 2016.

Levels of Model Transparency

Simulatability

Understanding of the functioning of the **model**

- Can a human *easily* predict outputs?
- Can a human examine the model all at once?

Transparent model

Decomposability

Understanding at the level of **single components** (e.g. parameters)

Transparent Model Components

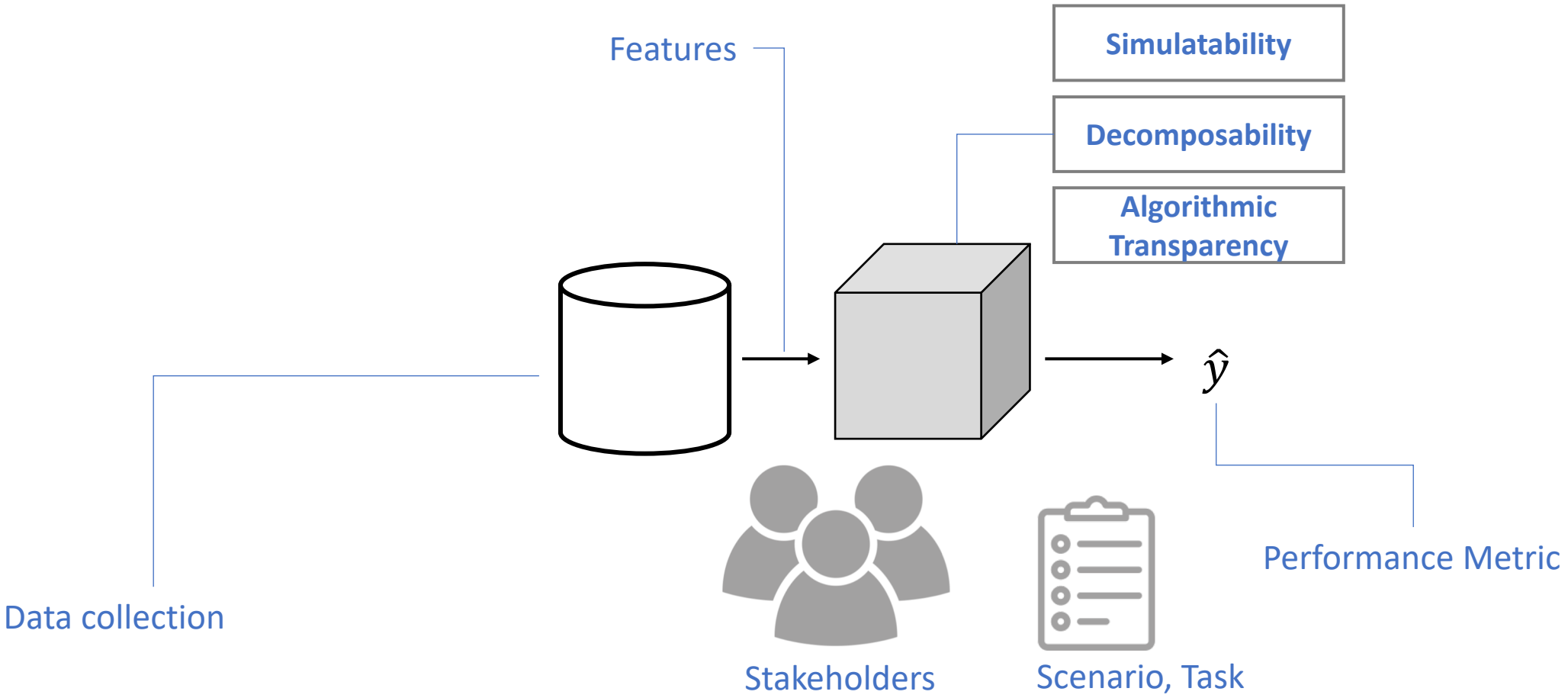
Algorithmic Transparency

Understanding at the level of **training algorithm**

Transparent Training Algorithm

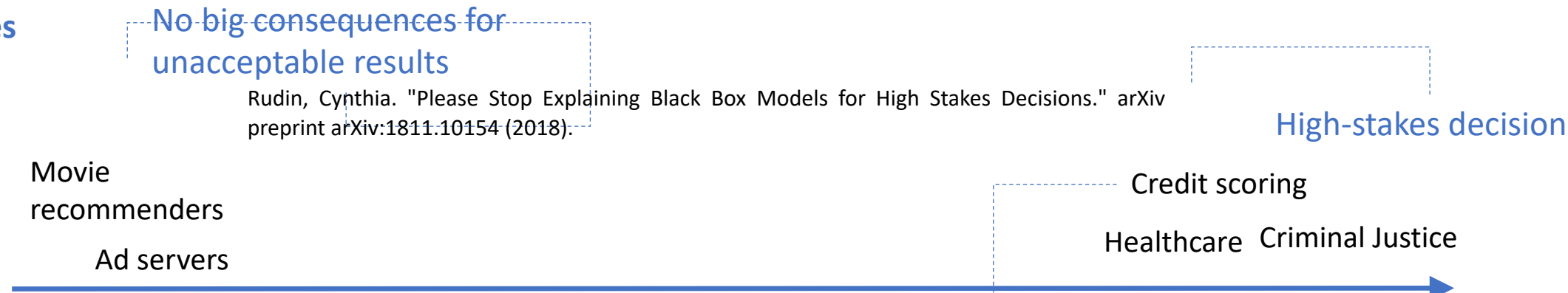
[Lipton 2016, Lepri et al. 2017, Mittelstadt et al. 2018, Weld and Bansal 2018]

Interpretability Goes Beyond the Model



Desire for Explainable AI **Must be Justified**

Interpretability comes at cost: Trade-off interpretability/predictive power



Freitas, Alex A. "Comprehensible classification models: a position paper." ACM SIGKDD explorations newsletter 15.1 (2014): 1-10.

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

Doshi-Velez, Finale, and Been Kim. "Towards a rigorous science of interpretable machine learning." arXiv preprint arXiv:1702.08608 (2017).

Lipton, Zachary C. "The mythos of model interpretability. Int. Conf." Machine Learning: Workshop on Human Interpretability in Machine Learning. 2016.

High-Stakes Scenarios Deserve Transparent Models

- Post-hoc explanations can be unreliable
- Design white-box, interpretable models straight away!
- (Or retro-fit approximate but interpretable models over complex ones)
- Problem: with thousands+ features DNNs perform better: post-hoc explanation the only way (?)

Rudin, Cynthia. "Please Stop Explaining Black Box Models for High Stakes Decisions." arXiv preprint arXiv:1811.10154 (2018).

Wachter, Sandra, Brent Mittelstadt, and Luciano Floridi. "Why a right to explanation of automated decision-making does not exist in the general data protection regulation." International Data Privacy Law 7.2 (2017): 76-99.

On Role of Data In XAI

Interpretable Data for Interpretable Models

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

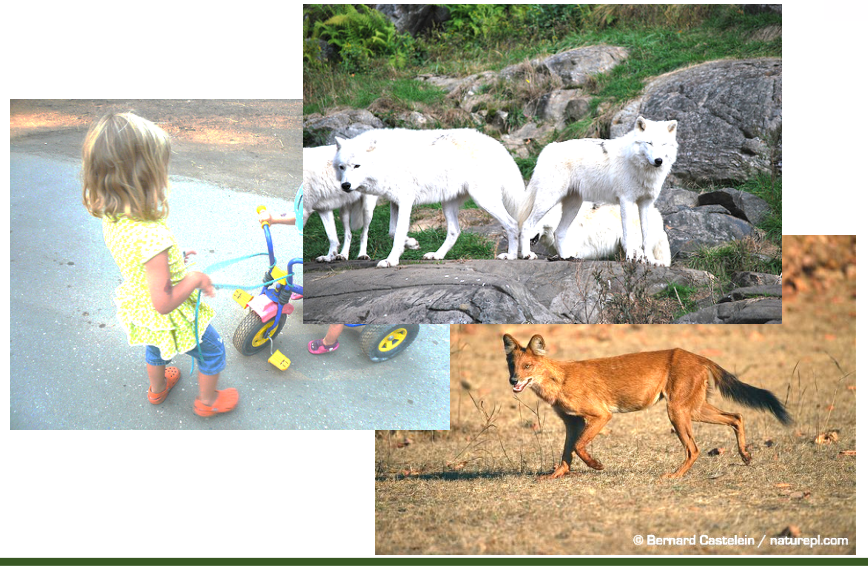
Field names

One row (4 fields)

2000 rows all told

Tabular

Images



Text

XAI Properties

(Some) Desired Properties of Explainable AI Systems (1)

- **Informativeness:** to which extend the model / prediction can be of use
- ***Interpretability*** *(or* comprehensibility*)*: to which extent the model and/or its predictions are human understandable. Is measured with the *complexity* of the model.
- ***Fidelity***: to which extent the model imitate a black-box predictor.
- ***Accuracy***: to which extent the model predicts unseen instances.

Alex A. Freitas. 2014. ***Comprehensible classification models: A position paper***. ACM SIGKDD Explor. Newslett.

Doshi-Velez, Finale, and Been Kim. "Towards a rigorous science of interpretable machine learning." arXiv preprint arXiv:1702.08608 (2017).

Rudin, Cynthia. "Please Stop Explaining Black Box Models for High Stakes Decisions." arXiv preprint arXiv:1811.10154 (2018).

(Some) Desired Properties of Explainable AI Systems (2)

- ***Fairness***: the model guarantees the protection of groups against discrimination.
- ***Privacy***: the model does not reveal sensitive information about people.
- ***Respect Monotonicity***: the increase of the values of an attribute either increase or decrease in a monotonic way the probability of a record of being member of a class.
- ***Usability***: an interactive and queryable explanation is more usable than a textual and fixed explanation.
- **Low cognitive load**: explanation should easy to understand

Andrea Romei and Salvatore Ruggieri. 2014. ***A multidisciplinary survey on discrimination analysis***. Knowl. Eng.

Yousra Abdul Alsahib S. Aldeen, Mazleena Salleh, and Mohammad Abdur Razzaque. 2015. ***A comprehensive review on privacy preserving data mining***. SpringerPlus .

Alex A. Freitas. 2014. ***Comprehensible classification models: A position paper***. ACM SIGKDD Explor. Newslett.

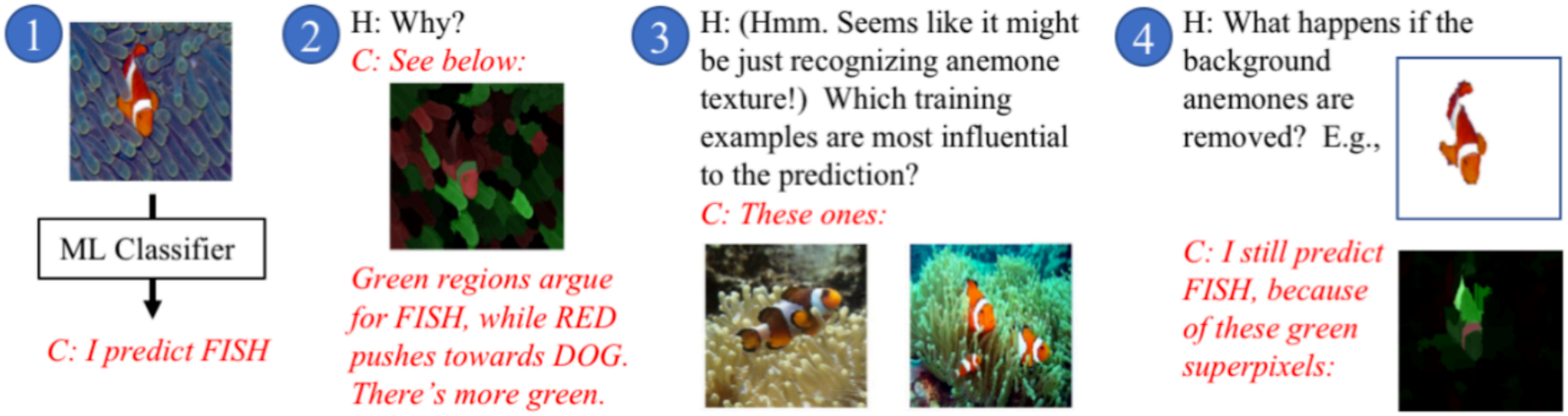
(Some) Desired Properties of Explainable AI Systems (3)

- ***Reliability and Robustness***: the interpretable model should maintain high levels of performance independently from small variations of the parameters or of the input data.
- ***Non-misleading:*** the interpretation sticks to the models, and do not hallucinate on behavior
- ***Causality:*** controlled changes in the input due to a perturbation should affect the model behavior.
- ***Scalability:*** the interpretable model should be able to scale to large input data with large input spaces.
- ***Generality:*** the model should not require special training or restrictions.
- **Interactivity /Conversational:** explanation should be refined based on user profile, preference and experience

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

Mittelstadt, Brent, Chris Russell, and Sandra Wachter. "Explaining explanations in AI." arXiv preprint arXiv:1811.01439 (2018).

Explanation as *System-Human Conversation*



- 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).

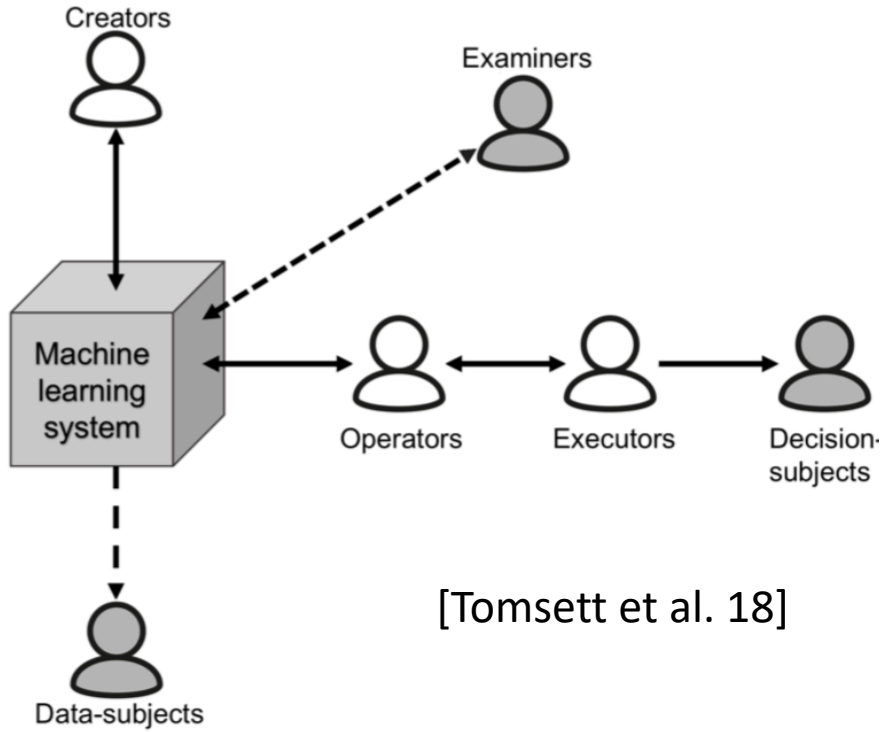
What about the Users?

Role-based Interpretability

"~~Is the system interpretable~~?" → "*To whom* is the system interpretable?"

No Universally Interpretable Model!

- **End users** "Am I being treated fairly?"
  "Can I contest the decision?"
  "What could I do differently to get a positive outcome?"
- **Engineers, data scientists**: "Is my system working as designed?"
- **Regulators** " Is it compliant?"
- **C-suite**



[Tomsett et al. 18]

An ideal explainer should model the *user background.*

[Tomsett et al. 2018, Weld and Bansal 2018, Poursabzi-Sangdeh 2018, Mittelstadt et al. 2019]

Designing Explanations is Task-Related

- Interpretability is always scenario-dependent!

  W*hat does interpretability mean* in a specific context? Ask the experts!

- What is the ultimate goal of the explanation in that specific **context,** for that specific **task**?
- How incomplete is the problem formulation?
- Time constraints
- Which user expertise?

Lipton, Zachary C. "The mythos of model interpretability. Int. Conf." Machine Learning: Workshop on Human Interpretability in Machine Learning. 2016.

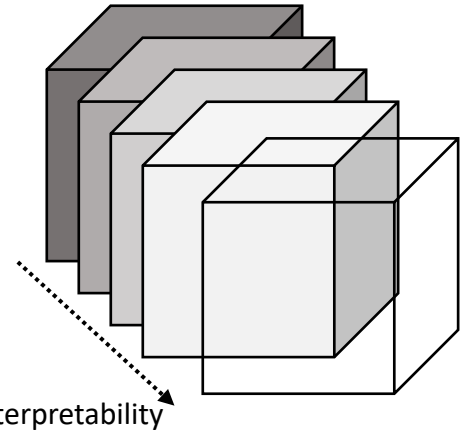
Doshi-Velez, Finale, and Been Kim. "Towards a rigorous science of interpretable machine learning." arXiv preprint arXiv:1702.08608 (2017).

Rudin, Cynthia. "Please Stop Explaining Black Box Models for High Stakes Decisions." arXiv preprint arXiv:1811.10154 (2018).

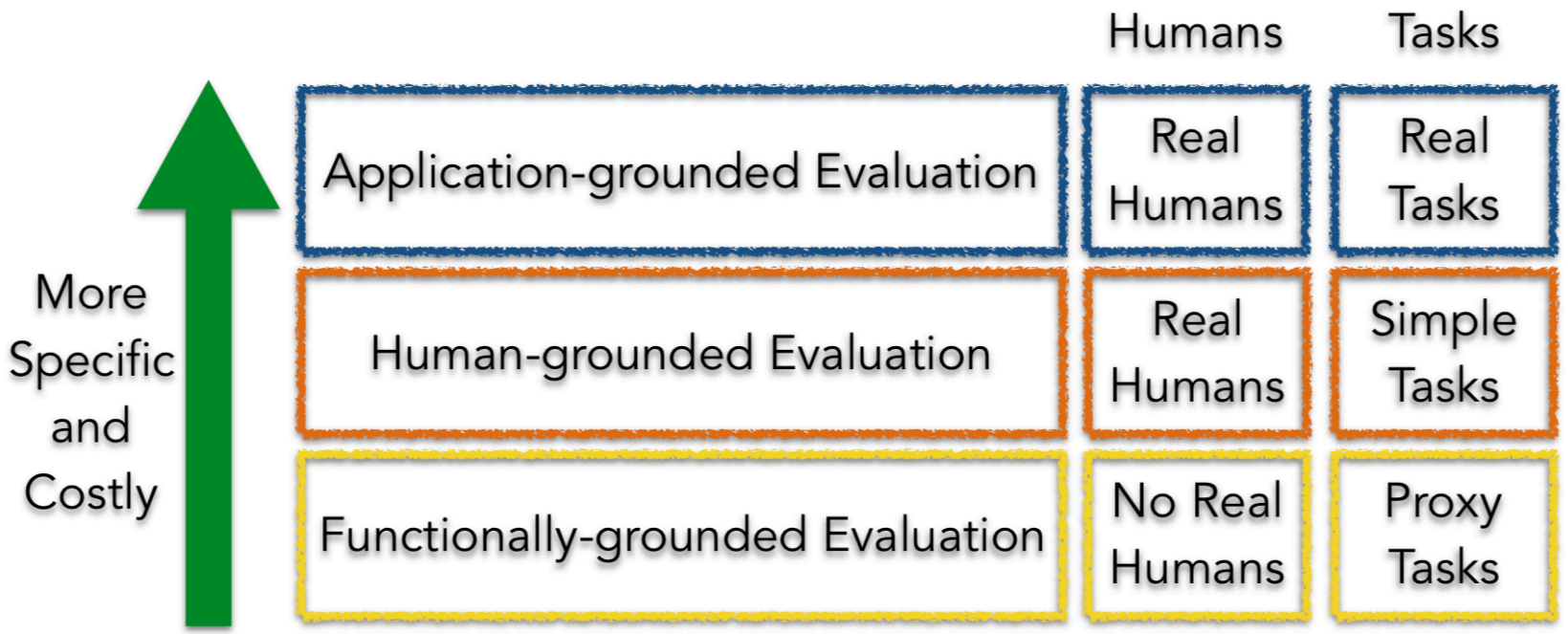
What about the Evaluation?

Evaluation: Interpretability as Latent Property

- Not directly measurable!
- Rely instead on *measurable outcomes*:
  - Any useful to individuals?
  - Can user estimate what a model will predict?
  - How much do humans follow predictions?
  - How well can people detect a mistake?
- No established benchmarks
- How to rank interpretable models? Different degrees of interpretability?



Evaluation Approaches



[Doshi-Velez and Kim 2017]

Human-Independent Metrics: Size

- Size is over-simplistic [Freitas 14]
  - E.g.: # nodes in a decision tree, size of a local explanation
  - Humans can handle at most 7±2 symbols [Miller1956, Rudin2018]
  - Size does not capture *semantics* of the model
  - Extreme simplicity insufficient! e.g. medical experts and larger models, [Freitas 2014]
  - What does *too large* mean?

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

Human-based Evaluation is Essential

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]

Human-based Evaluation for Feature Attribution-based Approaches

Have humans review attributions and/or compare them to (human provided) groundtruth on “feature importance”

Pros:

- Helps assess if attributions are human-intelligible
- Helps increase trust in the attribution method

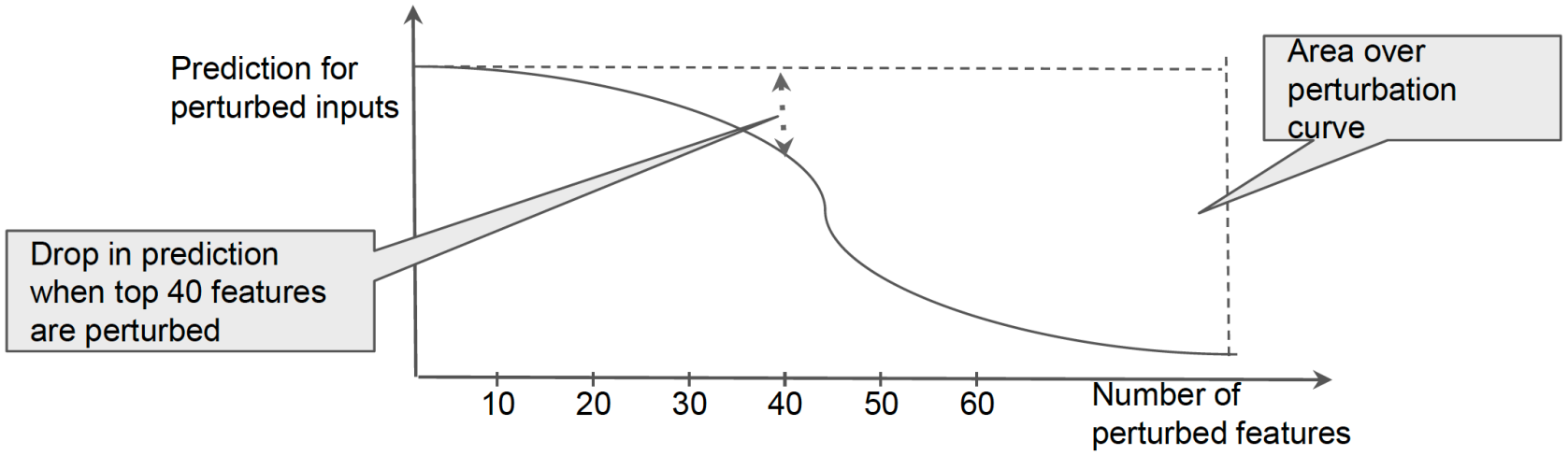
Cons:

- Attributions may appear incorrect because model reasons differently
- **Confirmation bias**

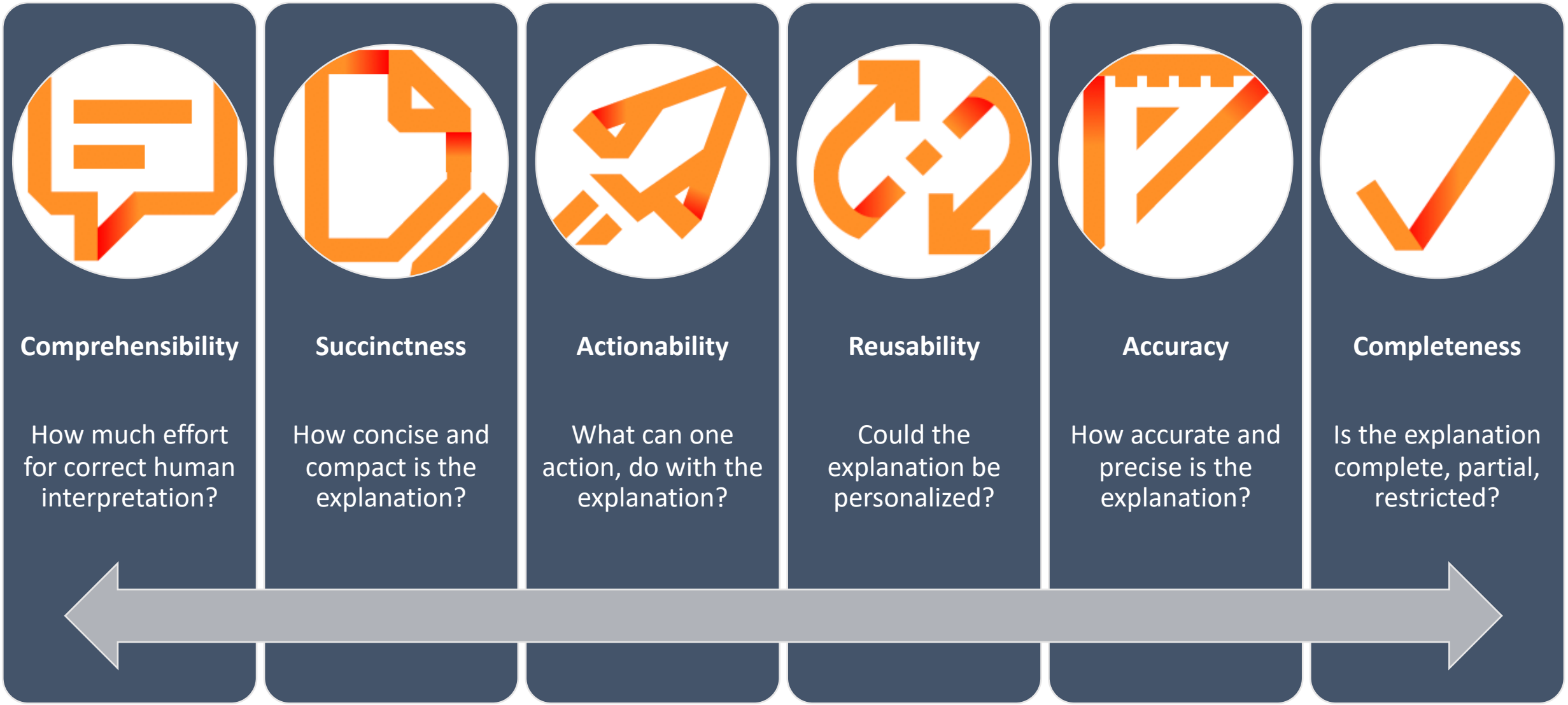
Perturbation-based Evaluation for Feature Attribution-based Approaches

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



XAI: One Objective, Many Metrics



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

Open Challenges

- More formal studies on interpretability
- *Rigorous, agreed upon* evaluation protocols
- More work on transparent design
- Human involvement (e.g. better interactive, “social” explanations) [Miller 2017]
- Define industry standards (e.g. AI Service Factsheet [Hind et al. 2018)]
- Improve existing legislation
  - “Right to explanation” vs “right to be informed” [Wachter et al. 2017]
  - Legislation & Explanations: How accurate ? How complete? How faithful to the model? [Rudin 2018]

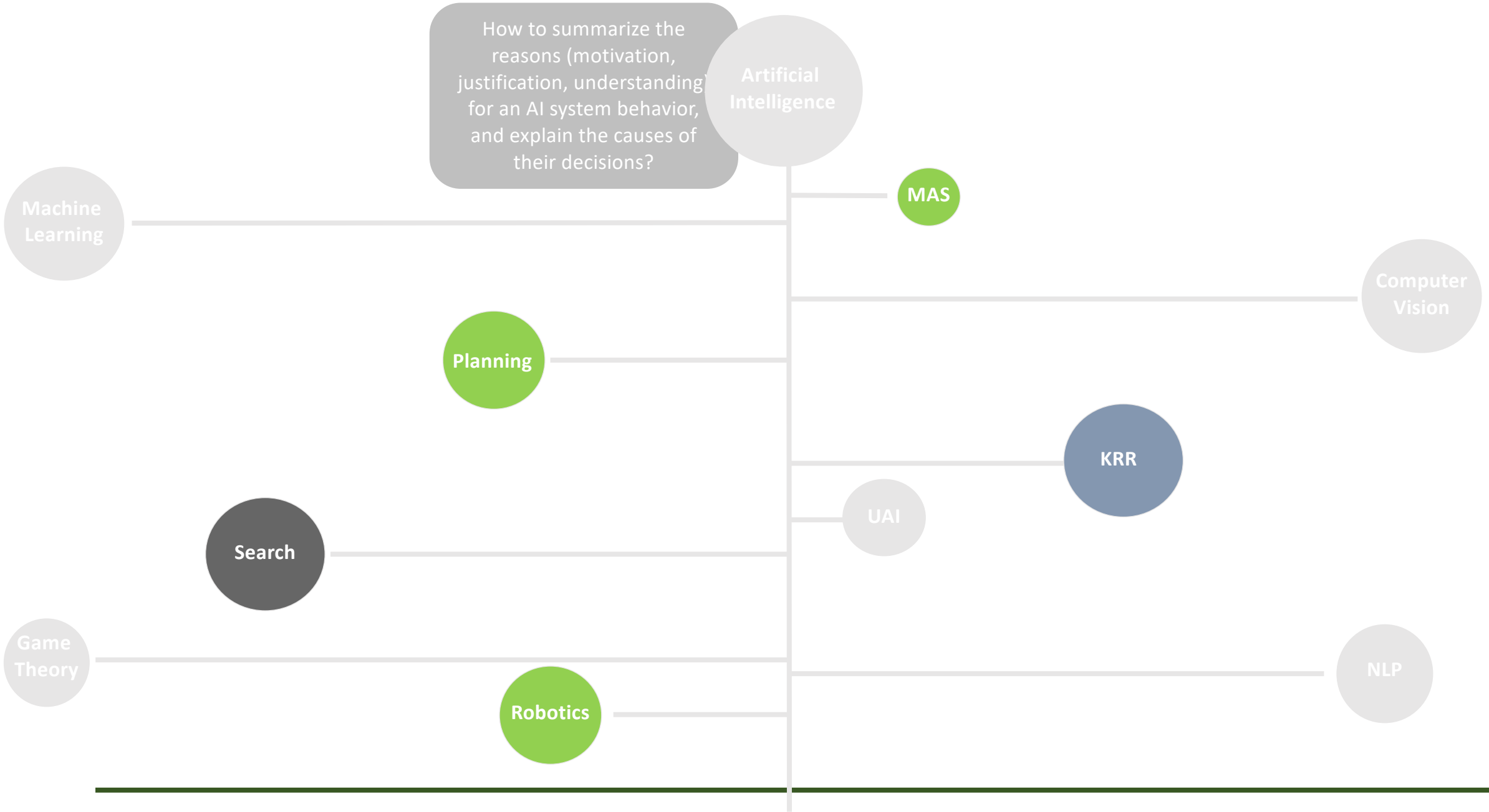
tl;dr

- Explanations and interpretability are required for better human trust, system debug, and legal compliance.
- No monolithic, agreed upon definition of Explainable AI
- Adoption spans multiple AI fields
- Explainability, interpretability come at a cost
- Design with humans and task in mind
- Human-based evaluation is essential

XAI in AI

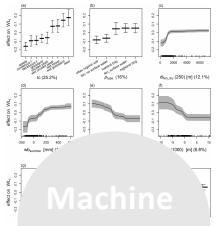
XAI: One Objective, Many 'AI's, Many Definitions, Many Approaches

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

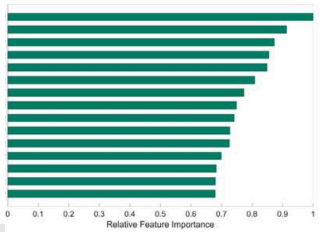


XAI: One Objective, Many 'AI's, Many Definitions, Many Approaches

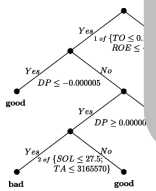
Dependency Plot



Feature Importance



Surrogate Model



Machine Learning

Which features are responsible of classification?

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

Artificial Intelligence

MAS

Computer Vision

Planning

KRR

UAI

Search

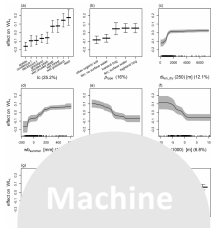
Game Theory

NLP

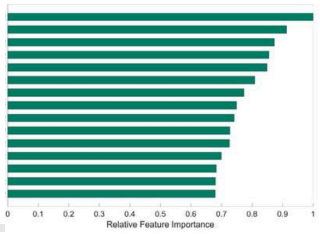
Robotics

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

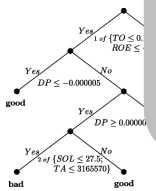
Dependency Plot



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?

Planning

Search

Game Theory

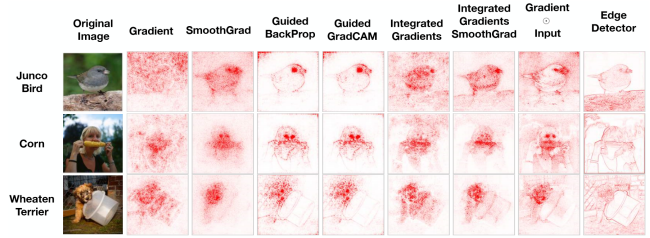
Robotics

MAS

KRR

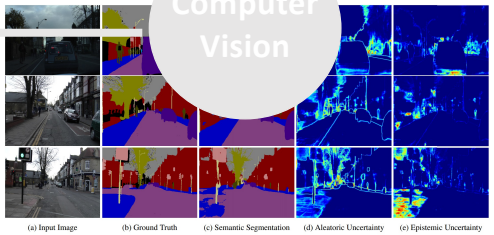
UAI

Saliency Map



Which complex features are responsible of classification?

Computer Vision

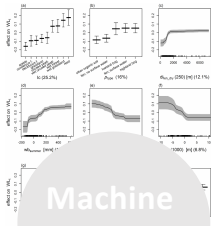


Uncertainty Map

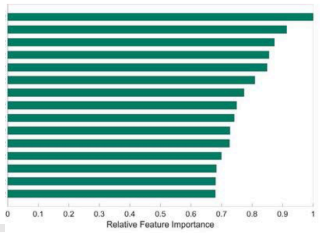
NLP

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

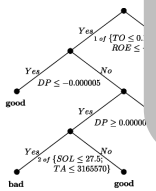
Dependency Plot



Feature Importance



Surrogate Model



Machine Learning

Which features are responsible of classification?

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

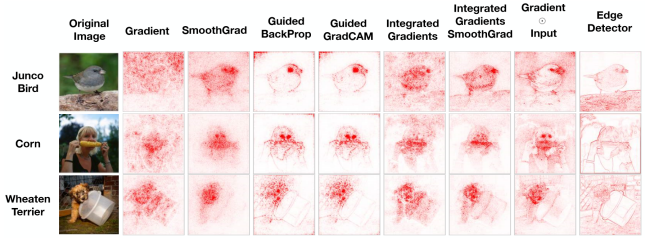
Artificial Intelligence

Strategy Summarization

MAS

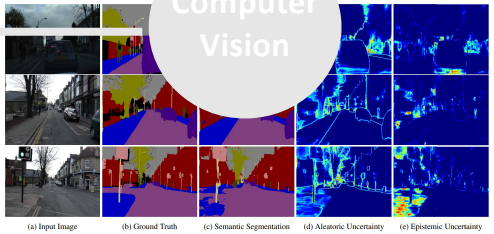
- Which agent strategy & plan ?
- Which player contributes most?
- Why such a conversational flow?

Saliency Map



Which complex features are responsible of classification?

Computer Vision



Uncertainty Map

Planning

KRR

UAI

Search

Game Theory

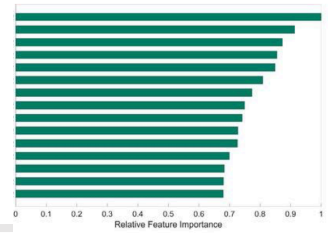
Robotics

NLP

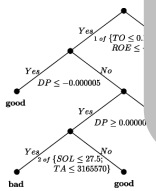
XAI: One Objective, Many 'AI's, Many Definitions, Many Approaches

Dependency Plot

Feature Importance



Surrogate Model



Machine Learning

Which features are responsible of classification?

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

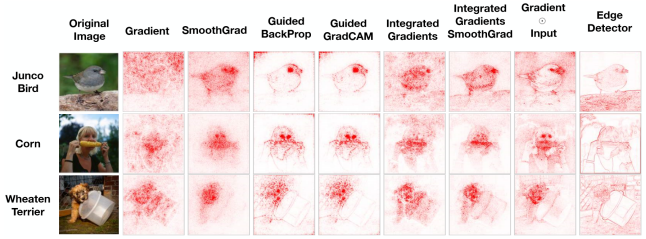
Artificial Intelligence

Strategy Summarization

MAS

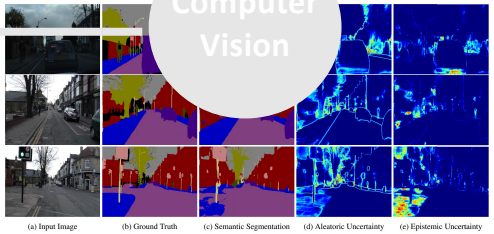
- Which agent strategy & plan ?
- Which player contributes most?
- Why such a conversational flow?

Saliency Map



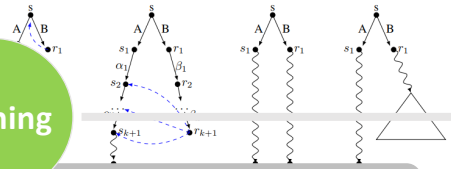
Which complex features are responsible of classification?

Computer Vision



Uncertainty Map

Plan Refinement



Planning

Which actions are responsible of a plan?

KRR

UAI

Search

Game Theory

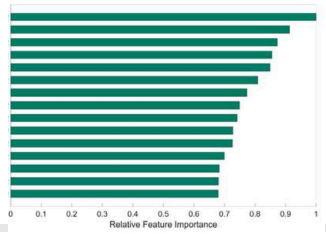
Robotics

NLP

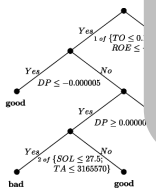
XAI: One Objective, Many 'AI's, Many Definitions, Many Approaches

Dependency Plot

Feature Importance



Surrogate Model



Machine Learning

Which features are responsible of classification?

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

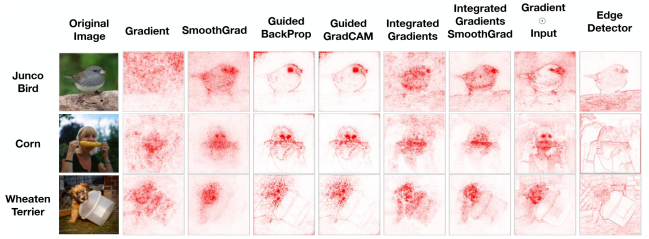
Artificial Intelligence

Strategy Summarization

MAS

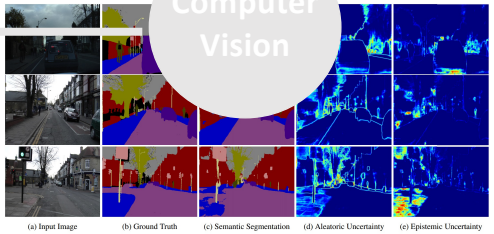
- Which agent strategy & plan ?
- Which player contributes most?
- Why such a conversational flow?

Saliency Map



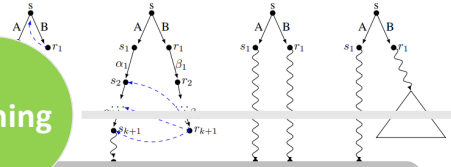
Which complex features are responsible of classification?

Computer Vision



Uncertainty Map

Plan Refinement



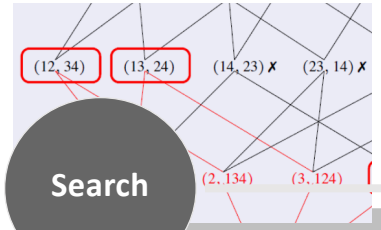
Planning

Which actions are responsible of a plan?

KRR

UAI

Conflicts Resolution



Search

Which constraints can be relaxed?

Game Theory

Robotics

NLP

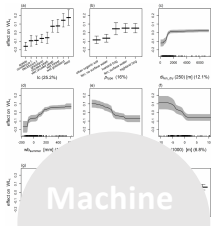
XAI: One Objective, Many 'AI's, Many Definitions, Many Approaches

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

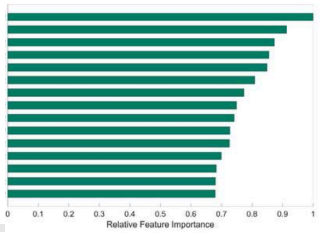
Artificial Intelligence

Machine Learning

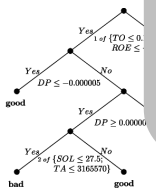
Dependency Plot



Feature Importance



Surrogate Model



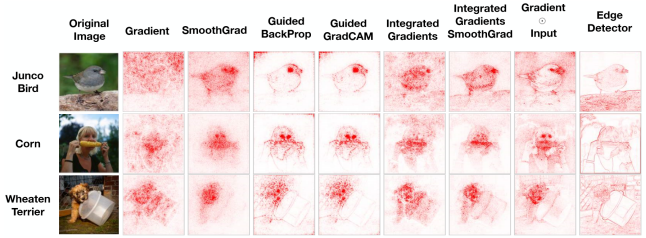
Which features are responsible of classification?

MAS

Strategy Summarization

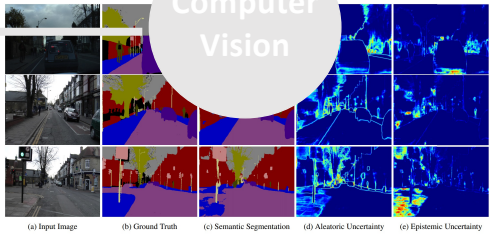
- Which agent strategy & plan ?
- Which player contributes most?
- Why such a conversational flow?

Saliency Map



Which complex features are responsible of classification?

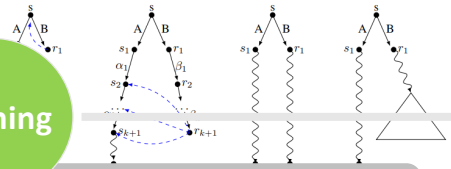
Computer Vision



Uncertainty Map

Planning

Plan Refinement



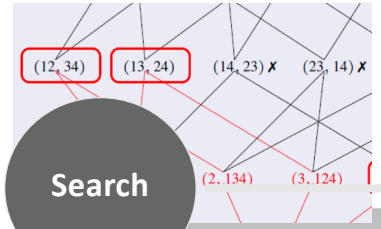
Which actions are responsible of a plan?

KRR

UAI

Conflicts Resolution

Search



Which constraints can be relaxed?

Game Theory

Which combination of features is optimal?

Robotics

NLP

Shapely Values

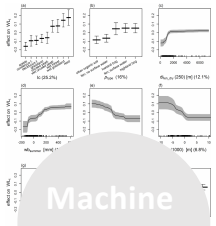
XAI: One Objective, Many 'AI's, Many Definitions, Many Approaches

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

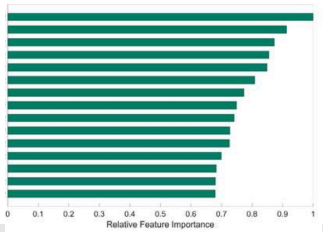
Artificial Intelligence

Machine Learning

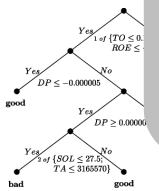
Dependency Plot



Feature Importance



Surrogate Model



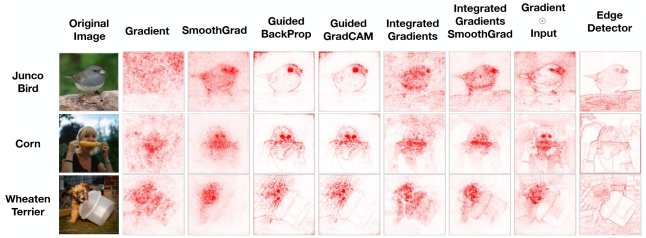
Which features are responsible of classification?

MAS

Strategy Summarization

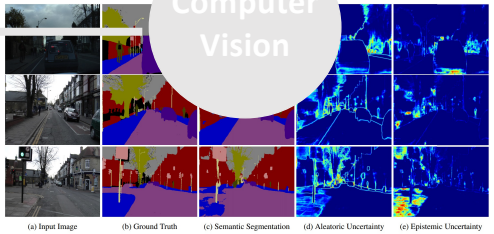
- Which agent strategy & plan ?
- Which player contributes most?
- Why such a conversational flow?

Saliency Map



Which complex features are responsible of classification?

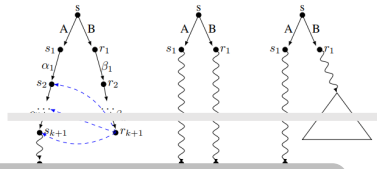
Computer Vision



Uncertainty Map

Planning

Plan Refinement



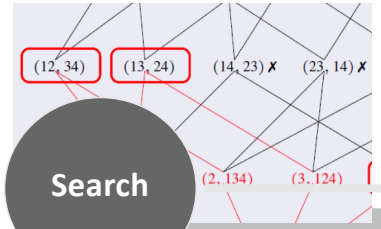
Which actions are responsible of a plan?

KRR

UAI

Conflicts Resolution

Search



Which constraints can be relaxed?

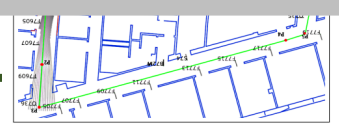
Game Theory

Which combination of features is optimal?

Shapely Values

Robotics

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



Narrative-based

NLP

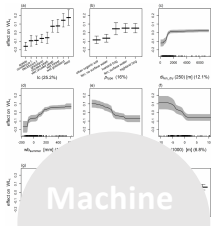
XAI: One Objective, Many 'AI's, Many Definitions, Many Approaches

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

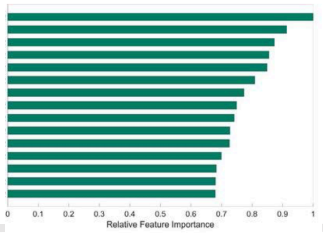
Artificial Intelligence

Machine Learning

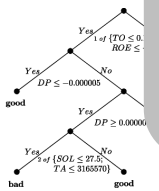
Dependency Plot



Feature Importance



Surrogate Model



Which features are responsible of classification?

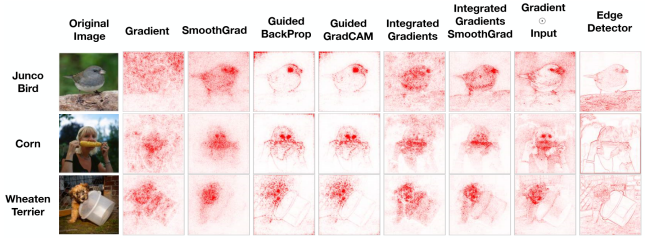
MAS

Strategy Summarization

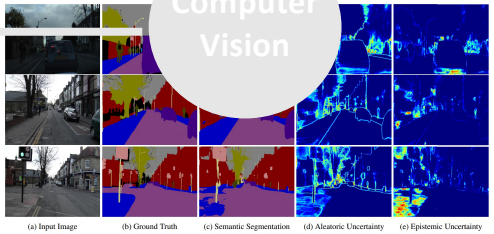
- Which agent strategy & plan ?
- Which player contributes most?
- Why such a conversational flow?

Computer Vision

Saliency Map



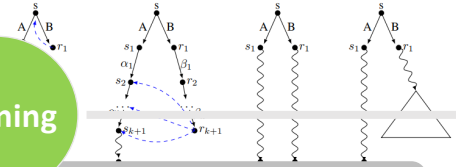
Which complex features are responsible of classification?



Uncertainty Map

Planning

Plan Refinement



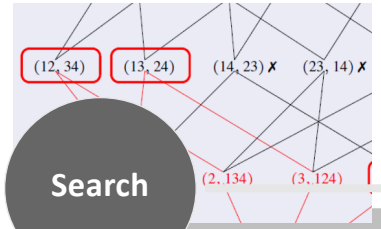
Which actions are responsible of a plan?

KRR

UAI

Conflicts Resolution

Search



Which constraints can be relaxed?

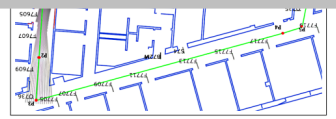
Game Theory

Which combination of features is optimal?

Shapely Values

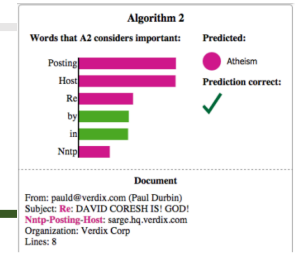
Robotics

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



Narrative-based

Machine Learning based



NLP

Which entity is responsible for classification?

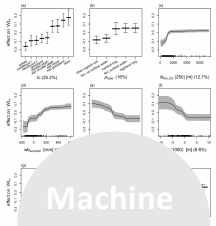
XAI: One Objective, Many 'AI's, Many Definitions, Many Approaches

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

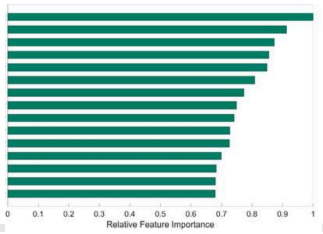
Artificial Intelligence

Machine Learning

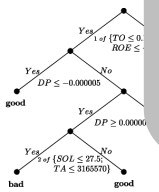
Dependency Plot



Feature Importance



Surrogate Model



Which features are responsible of classification?

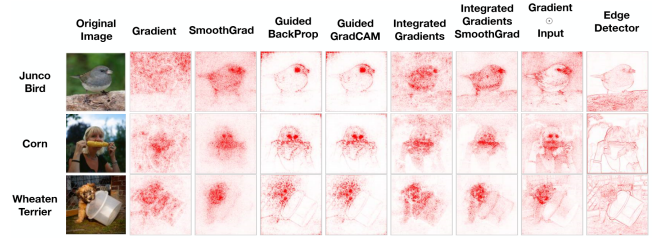
MAS

Strategy Summarization

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

Computer Vision

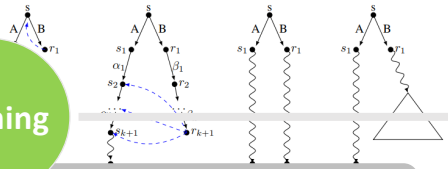
Saliency Map



Which complex features are responsible of classification?

Planning

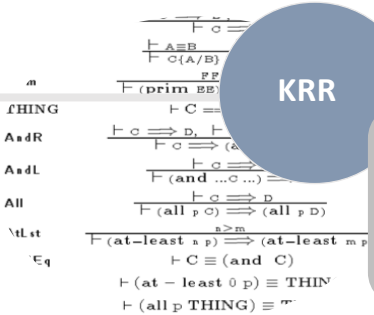
Plan Refinement



Which actions are responsible of a plan?

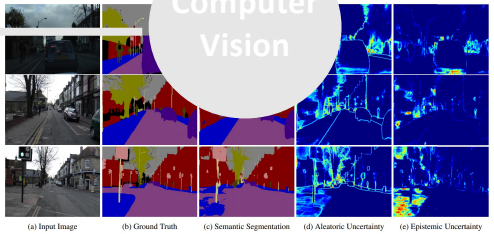
KRR

Diagnosis



Abduction

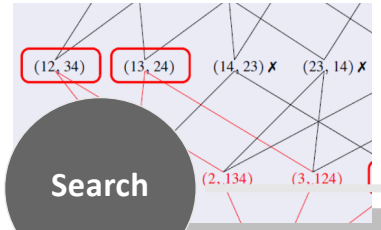
Uncertainty Map



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

Search

Conflicts Resolution



Which constraints can be relaxed?

UAI

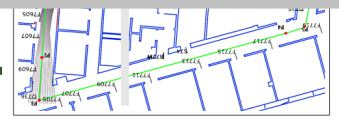
Game Theory

Which combination of features is optimal?

Shapely Values

Robotics

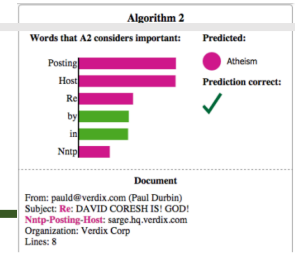
Which decisions, combination of multimodal decisions lead to an action?



Narrative-based

NLP

Machine Learning based



Which entity is responsible for classification?

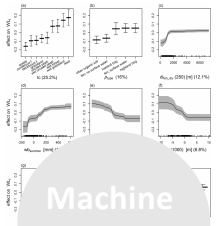
XAI: One Objective, Many 'AI's, Many Definitions, Many Approaches

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

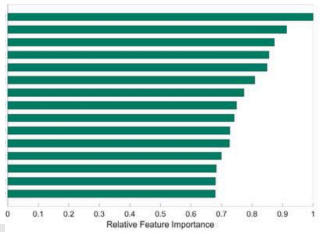
Artificial Intelligence

Machine Learning

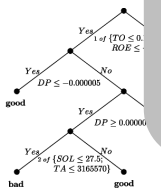
Dependency Plot



Feature Importance



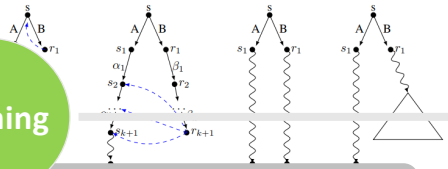
Surrogate Model



Which features are responsible of classification?

Planning

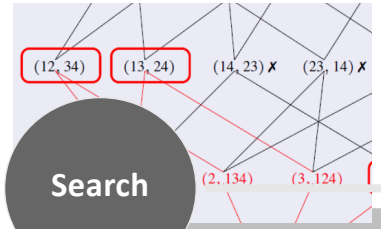
Plan Refinement



Which actions are responsible of a plan?

Search

Conflicts Resolution



Which constraints can be relaxed?

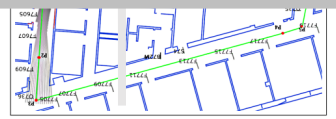
Game Theory

Which combination of features is optimal?

Shapely Values

Robotics

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



Narrative-based

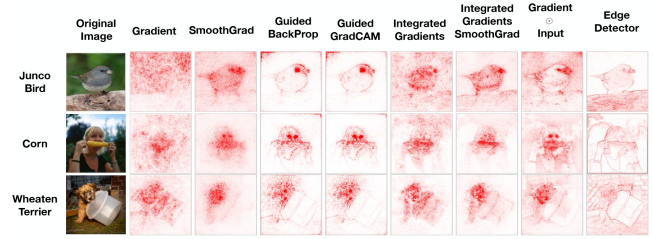
MAS

Strategy Summarization

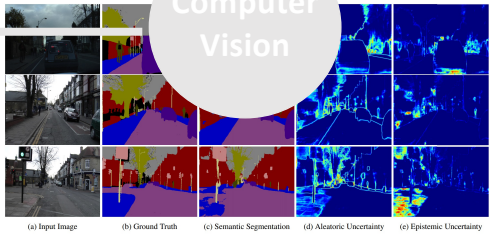
- Which agent strategy & plan ?
- Which player contributes most?
- Why such a conversational flow?

Computer Vision

Saliency Map



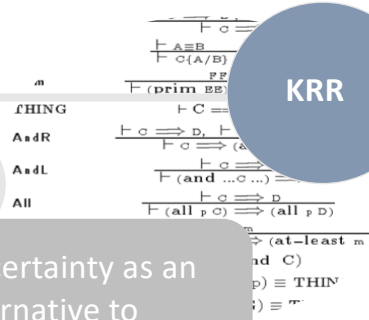
Which complex features are responsible of classification?



Uncertainty Map

KRR

Diagnosis



Abduction

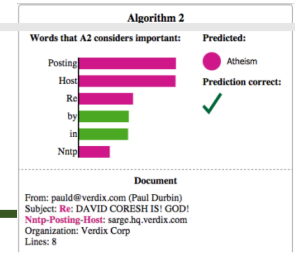
- Which axiom is responsible of inference (e.g., classification)?
- Abduction/Diagnostic: Find the **right** root causes (abduction)?

UAI

Uncertainty as an alternative to explanation

Machine Learning based

NLP



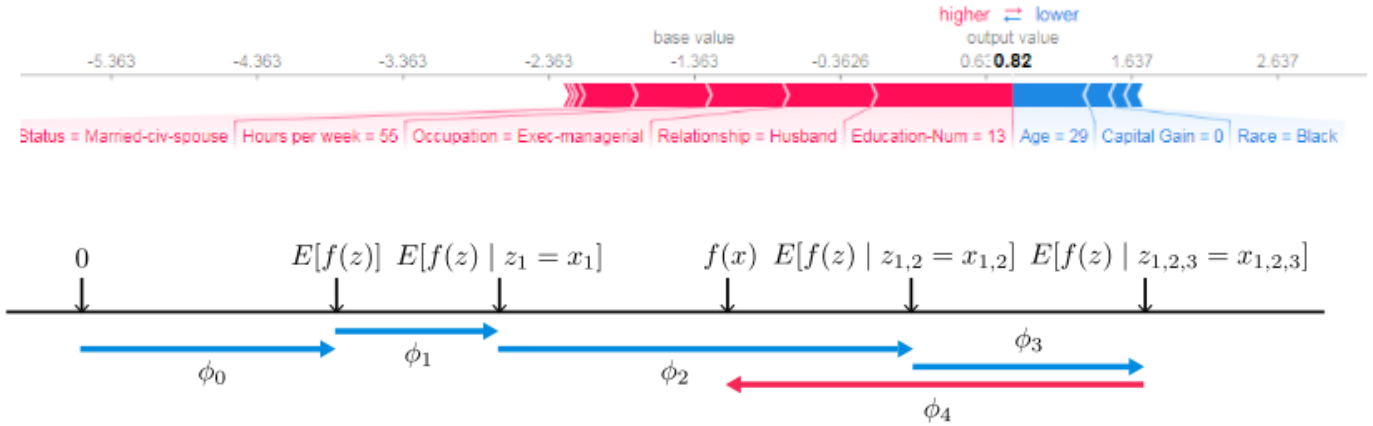
Which entity is responsible for classification?

Deep Dive on XAI in AI

(except Machine Learning)

Overview of explanation in different AI fields (1)

- Game Theory

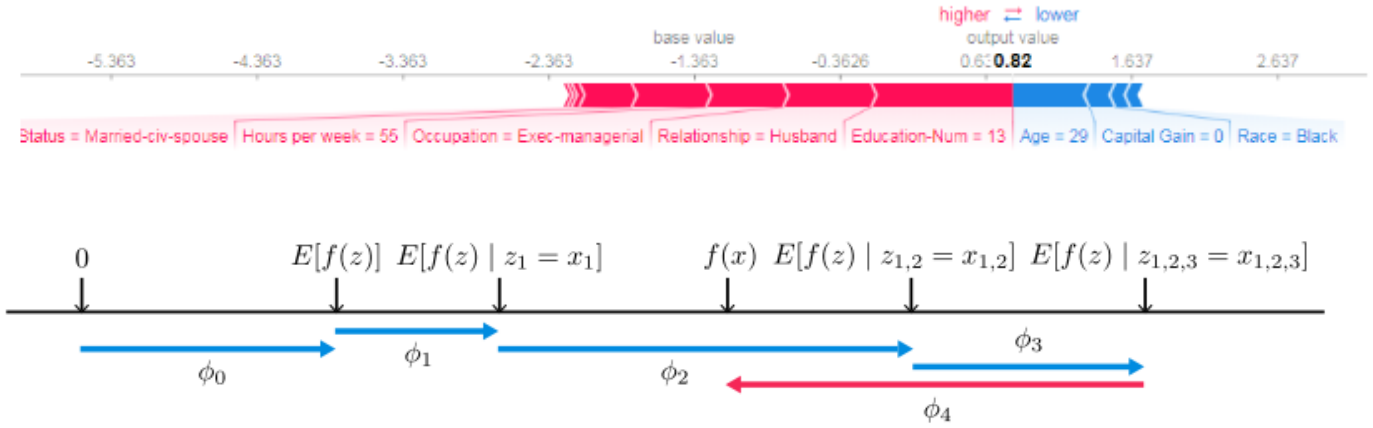


Shapley Additive Explanation

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

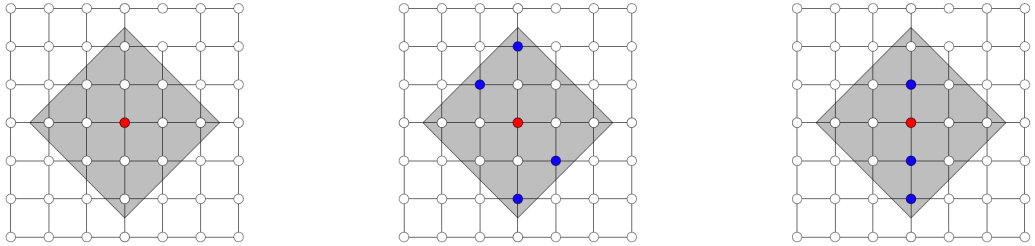
Overview of explanation in different AI fields (1)

- Game Theory



Shapley Additive Explanation

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

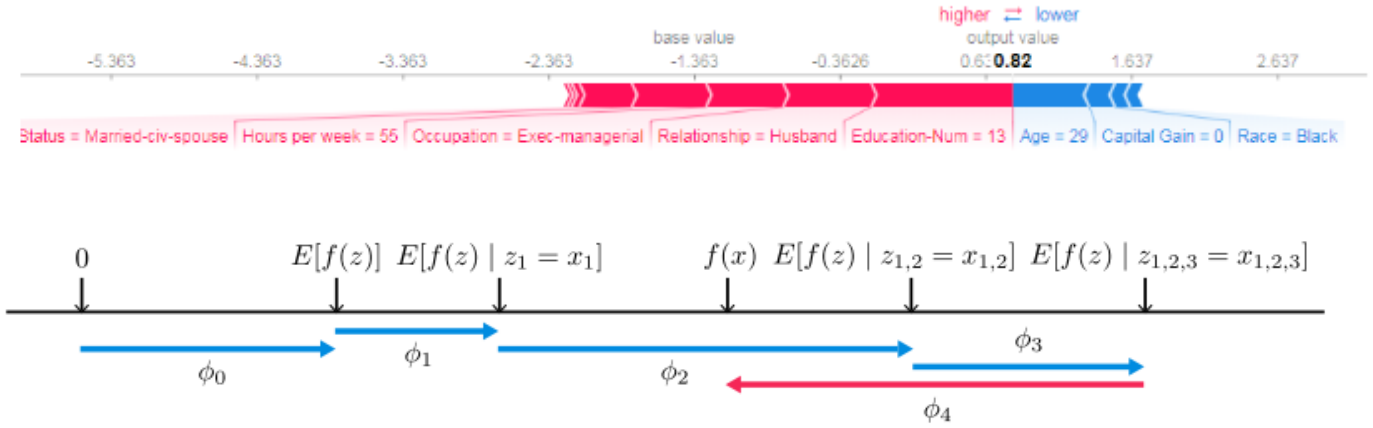


L-Shapley and C-Shapley (with graph structure)

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

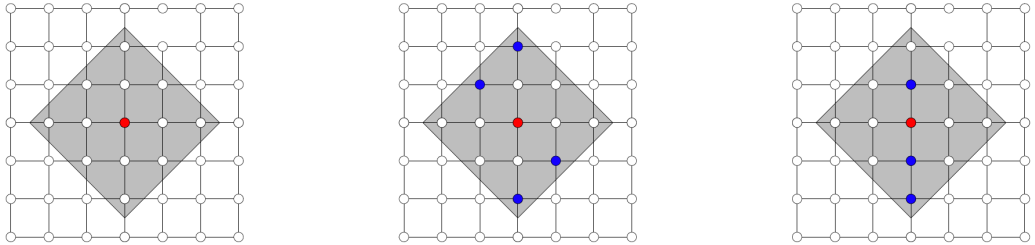
Overview of explanation in different AI fields (1)

- Game Theory



Shapley Additive Explanation

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



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

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

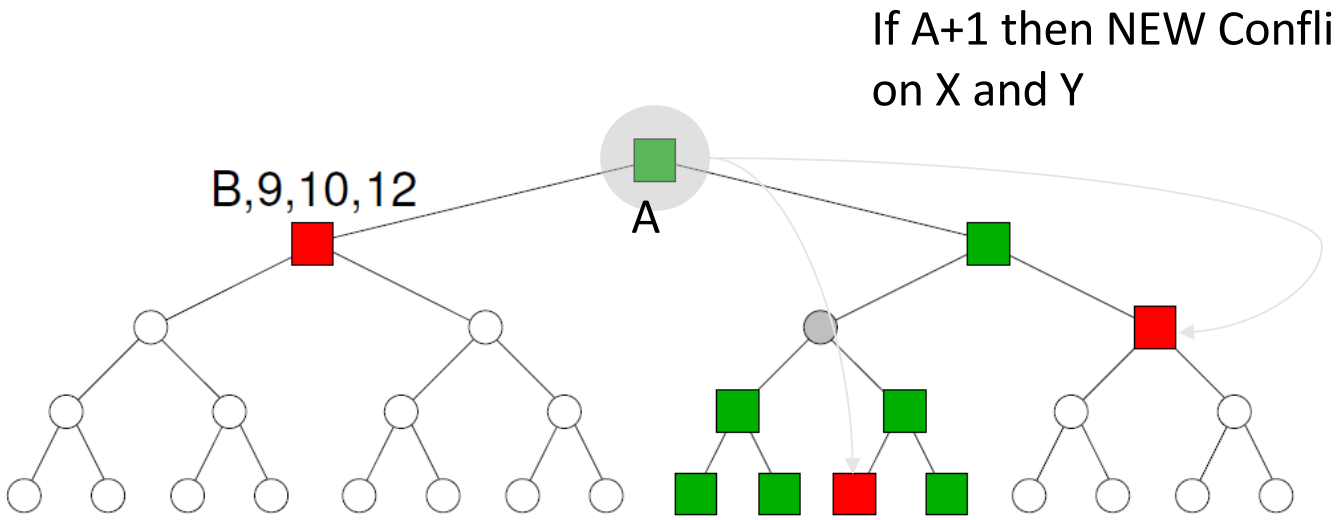
~ instancewise feature importance (causal influence)

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

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

Overview of explanation in different AI fields (2)

- Search and Constraint Satisfaction



Conflicts resolution

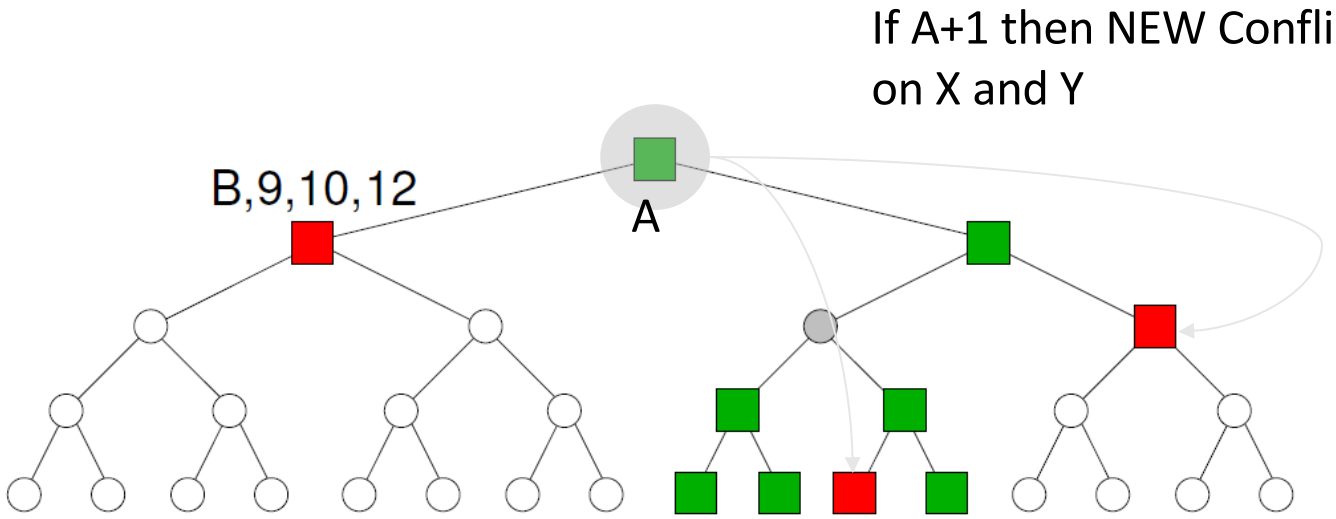
Barry O'Sullivan, Alexandre Papadopoulos, Boi Faltings, Pearl Pu: Representative Explanations for Over-Constrained Problems. AAAI 2007: 323-328

Robustness Computation

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

Overview of explanation in different AI fields (2)

- Search and Constraint Satisfaction

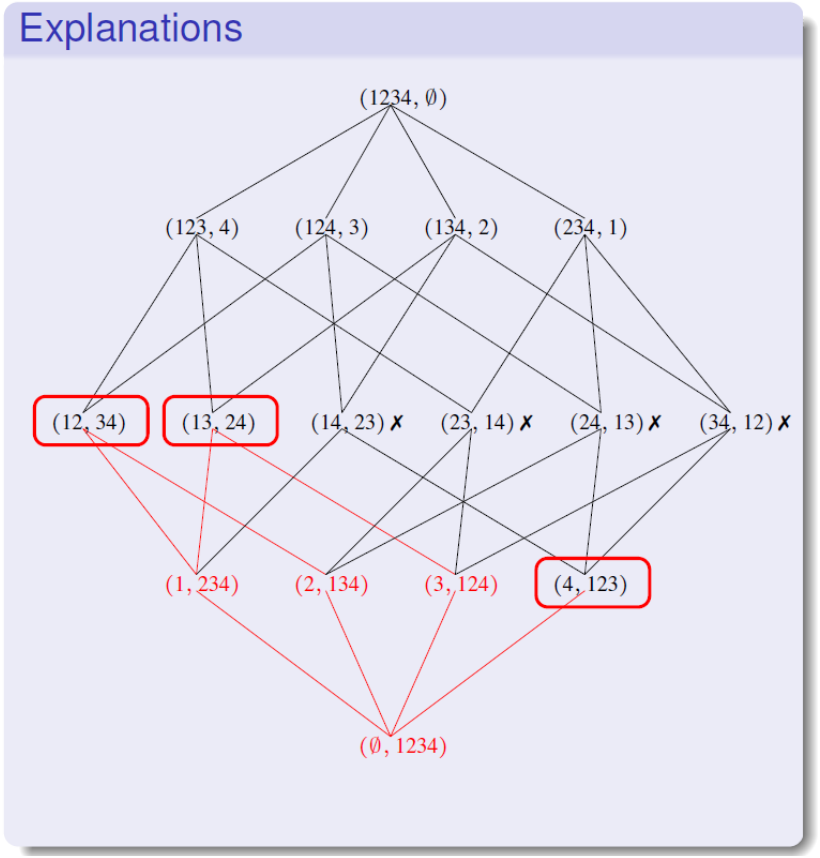


Conflicts resolution

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

Robustness Computation

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



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

A $\equiv$ (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

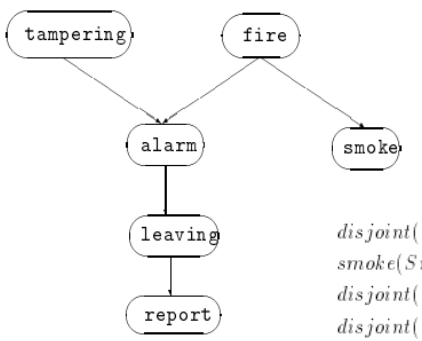
Ref	$\vdash C \Longrightarrow C$
Trans	$\frac{\vdash C \Longrightarrow D,\ \vdash D \Longrightarrow E}{\vdash C \Longrightarrow E}$
Eq	$\frac{\vdash A \equiv B \qquad \vdash C \Longrightarrow D}{\vdash C\{A/B\} \Longrightarrow D\{A/B\}}$
Prim	$\frac{FF \subset EE}{\vdash (\text{prim } EE) \Longrightarrow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

A $\equiv$ (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



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

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

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

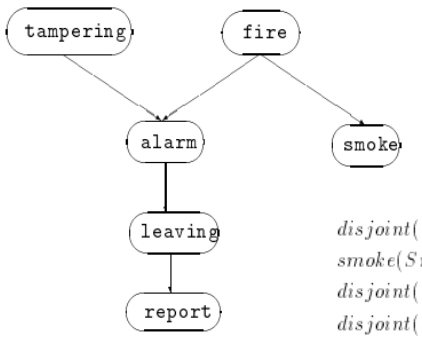
Ref	$\vdash C \Longrightarrow C$
Trans	$\dfrac{\vdash C \Longrightarrow D,\ \vdash D \Longrightarrow E}{\vdash C \Longrightarrow E}$
Eq	$\dfrac{\vdash A \equiv B \qquad \vdash C \Longrightarrow D}{\vdash C\{A/B\} \Longrightarrow D\{A/B\}}$
Prim	$\dfrac{FF \subset EE}{\vdash (\text{prim } EE) \Longrightarrow (\text{prim } FF)}$
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	$\vdash (\text{at-least } 0\ p) \equiv \text{THING}$
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

A $\equiv$ (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

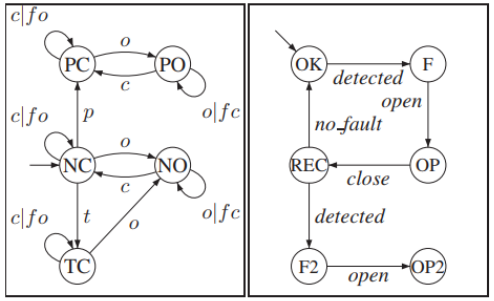


$$P(alarm|fire \wedge \neg tampering) = 0.99$$
$$P(alarm|\neg fire \wedge tampering) = 0.85$$
$$P(alarm|\neg fire \wedge \neg tampering) = 0.0001$$
$$P(leaving|alarm) = 0.88$$
$$P(leaving|\neg alarm) = 0.001$$
$$P(\text{report}|leaving) = 0.75$$
$$P(\text{report}|\neg leaving) = 0.01$$

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

Abduction Reasoning (in Bayesian Network)

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

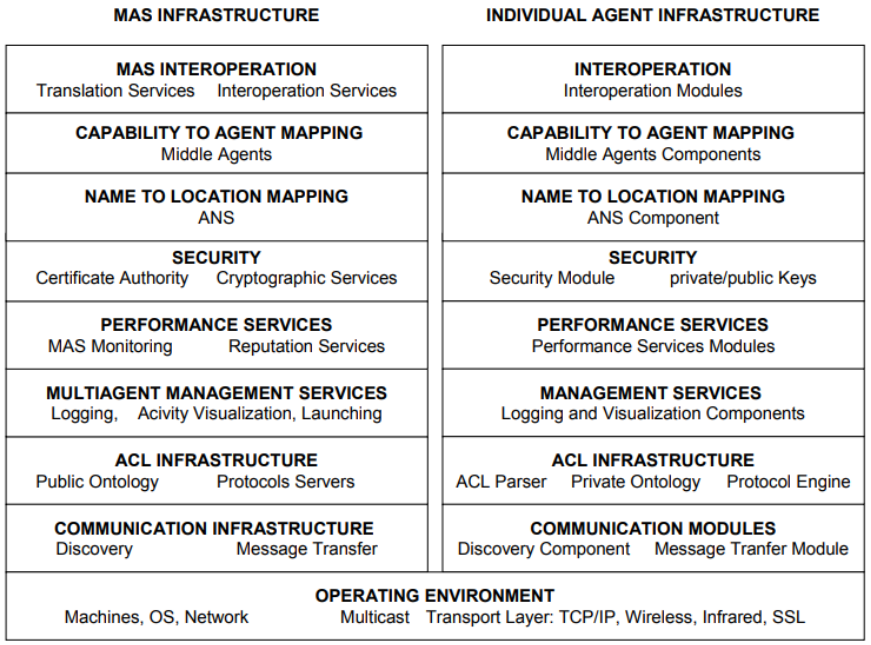


Diagnosis Inference

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)

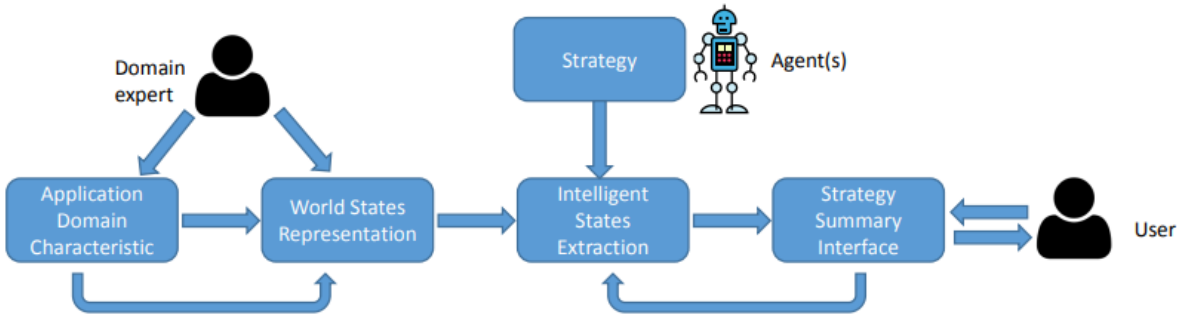
Overview of explanation in different AI fields (4)

- Multi-agent Systems

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

Explanation of Agent Conflicts & Harmful Interactions

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



Agent Strategy Summarization

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

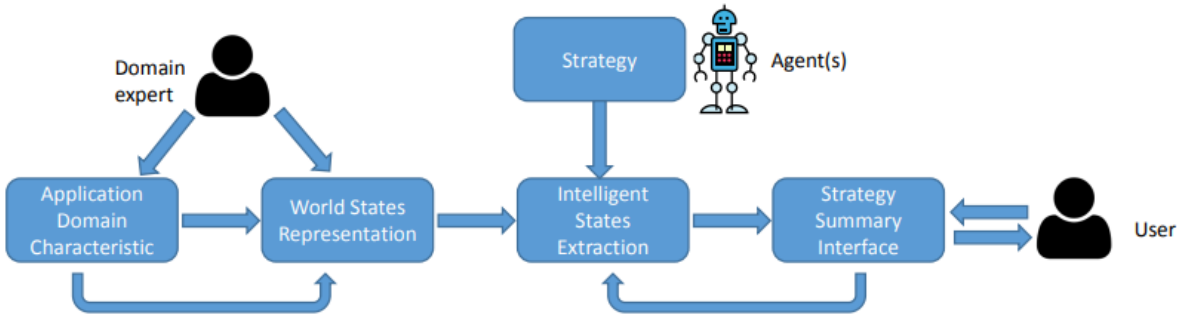
Overview of explanation in different AI fields (4)

- Multi-agent Systems

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

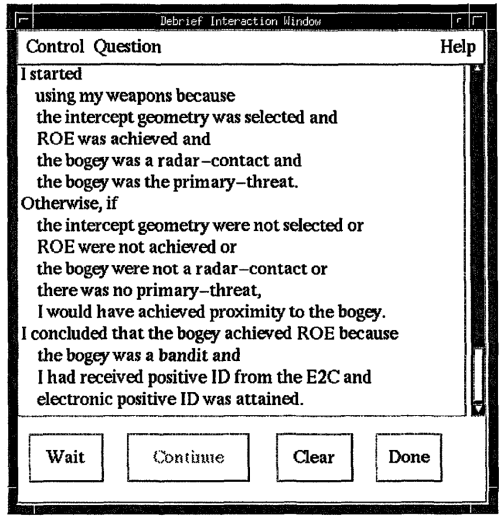
Explanation of Agent Conflicts & Harmful Interactions

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



Agent Strategy Summarization

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



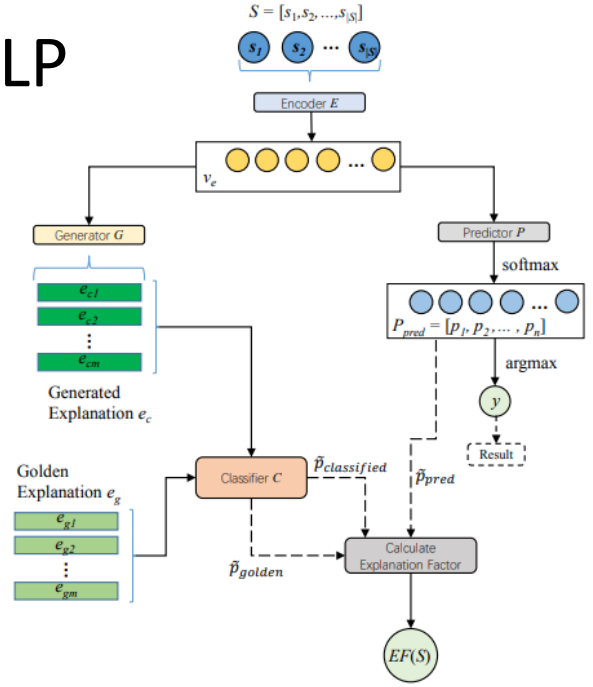
Explainable Agents

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

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

Overview of explanation in different AI fields (5)

- NLP



Fine-grained explanations are in the form of:

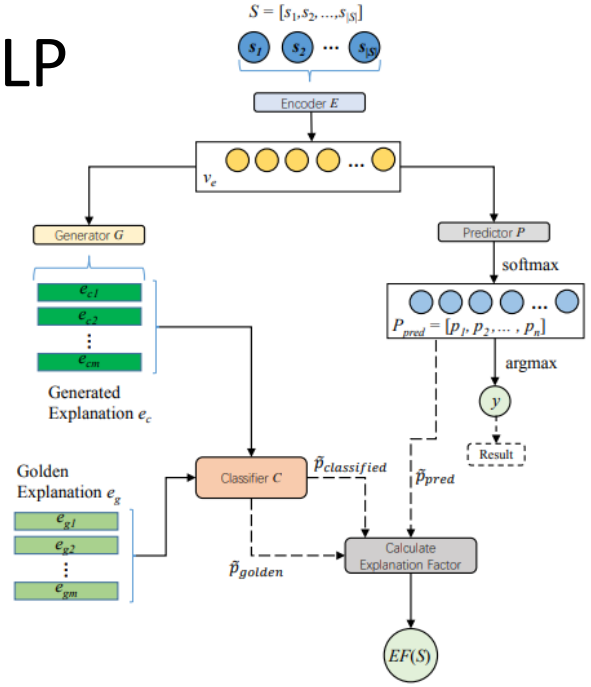
- texts in a real-world dataset;
- Numerical scores

Explainable NLP

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

Overview of explanation in different AI fields (5)

- NLP

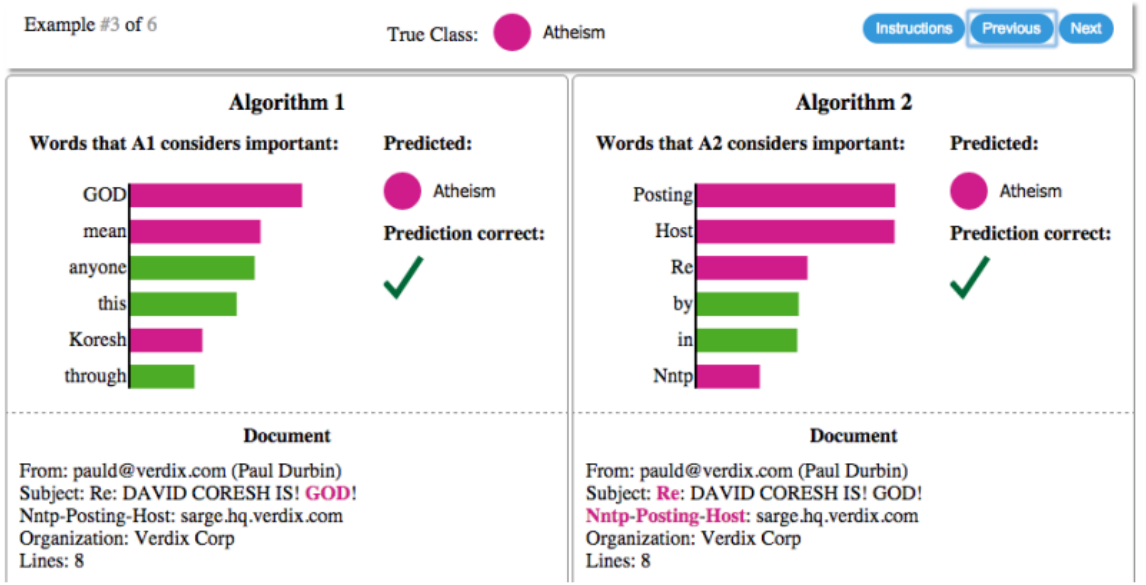


Explainable NLP

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

Fine-grained explanations are in the form of:

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

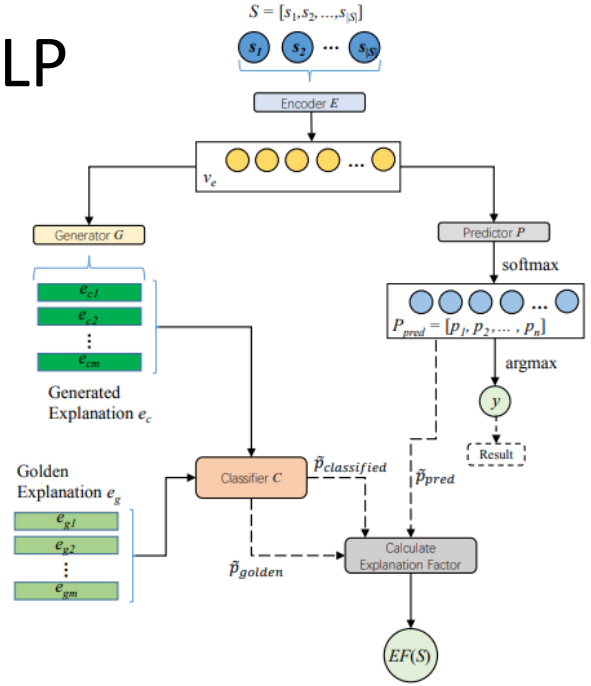


LIME for NLP

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

Overview of explanation in different AI fields (5)

- NLP

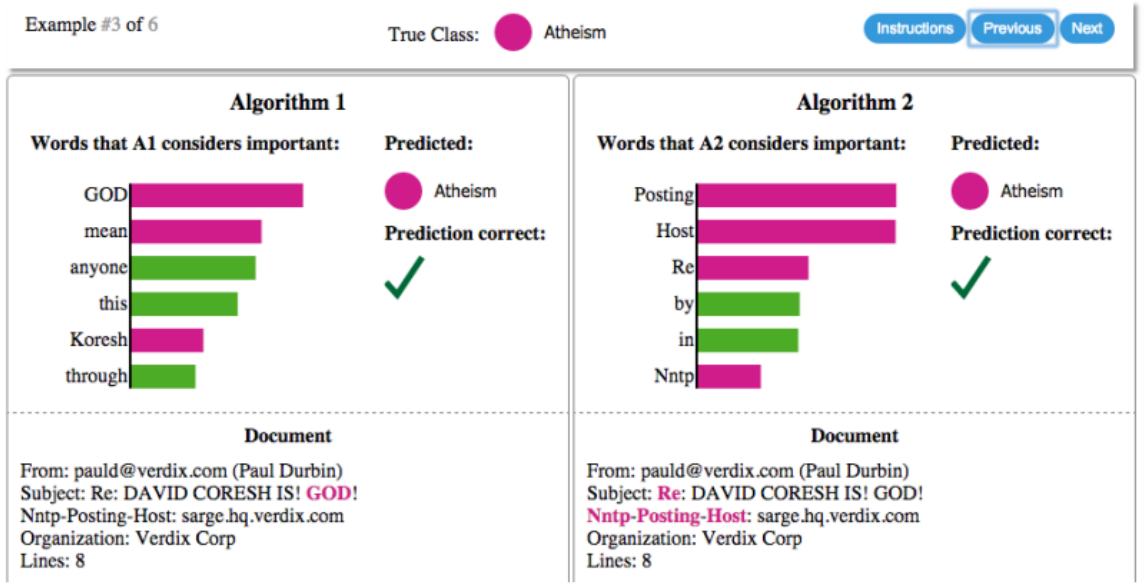


Fine-grained explanations are in the form of:

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

Explainable NLP

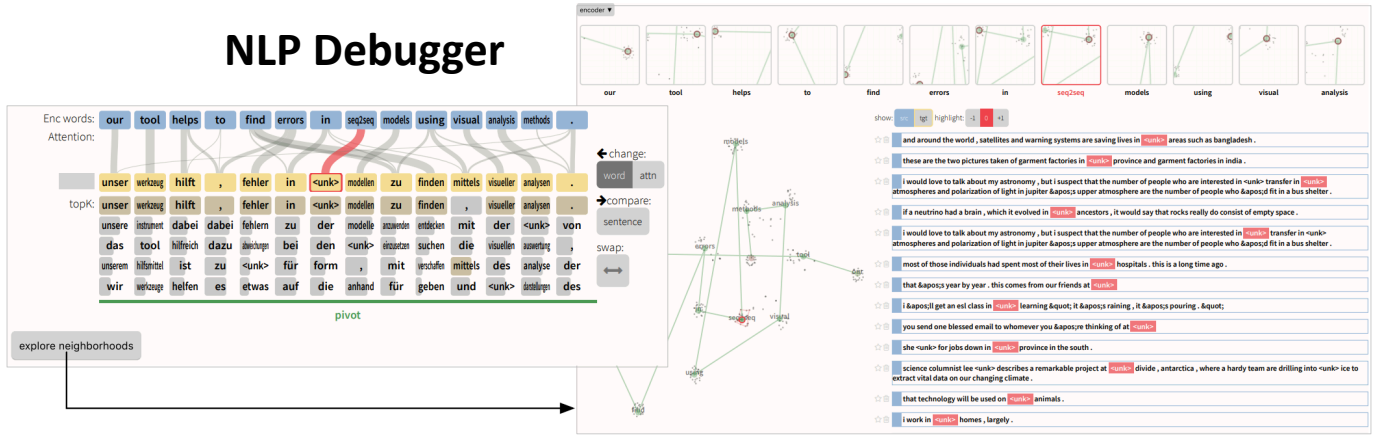
Hui Liu, Qingyu Yin, William Yang Wang: Towards Explainable NLP: A Generative Explanation Framework for Text Classification. CoRR abs/1811.00196 (2018)



LIME for NLP

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

NLP Debugger



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

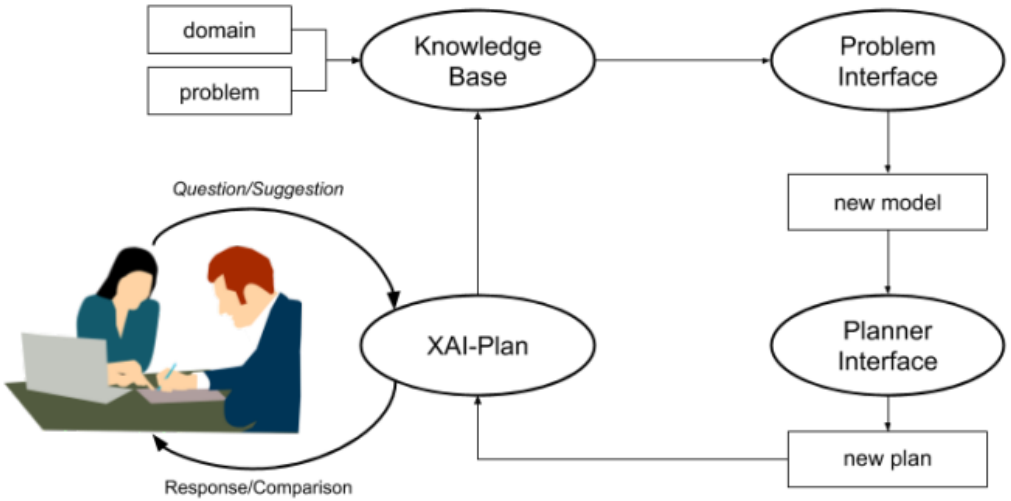
Hendrik Strobelt, Sebastian Gehrmann, Michael Behrisch, Adam Perer, Hanspeter Pfister, Alexander M. Rush: Seq2seq-Vis: A Visual Debugging Tool for Sequence-to-Sequence Models. IEEE Trans. Vis. Comput. Graph. 25(1): 353-363 (2019)

Overview of explanation in different AI fields (6)

- Planning and Scheduling

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

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



XAI Plan

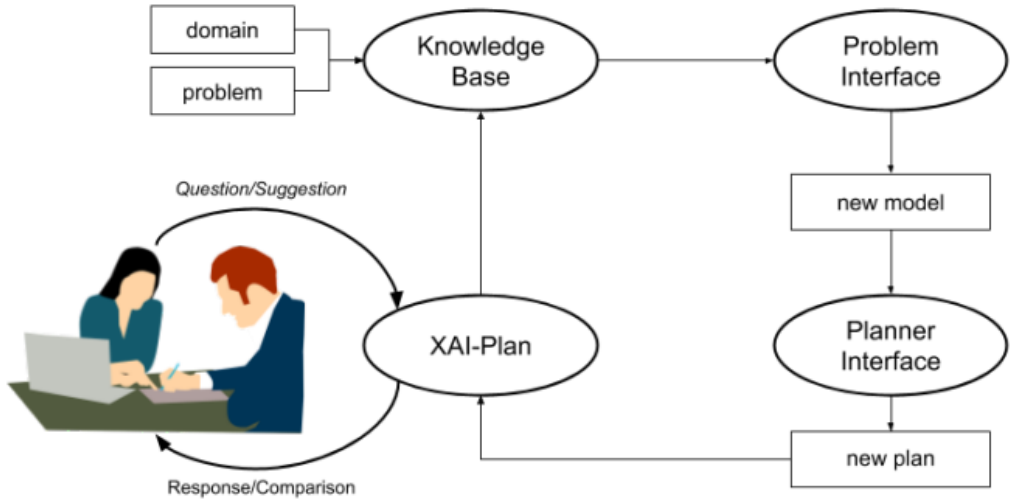
Rita Borgo, Michael Cashmore, Daniele Magazzeni: Towards Providing Explanations for AI Planner Decisions. CoRR abs/1810.06338 (2018)

Overview of explanation in different AI fields (6)

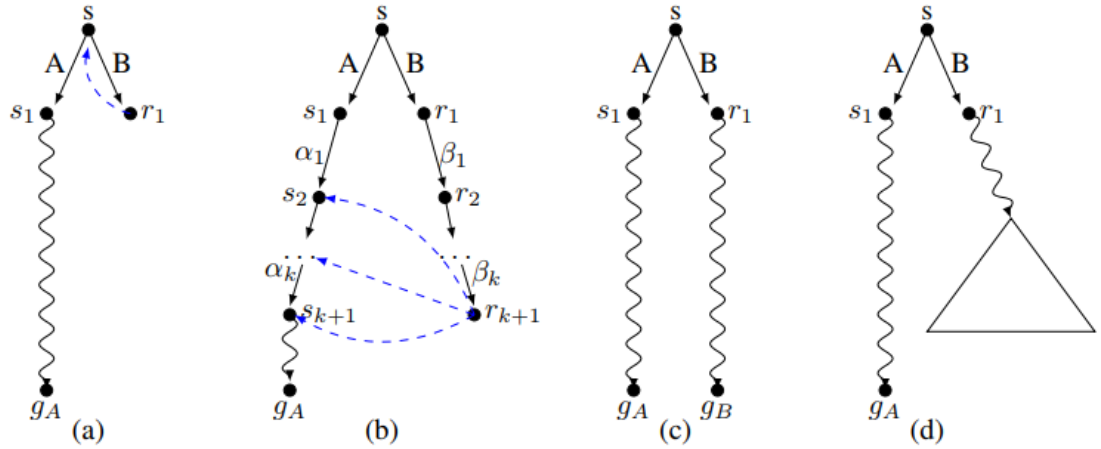
- Planning and Scheduling

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

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



XAI Plan



Human-in-the-loop Planning

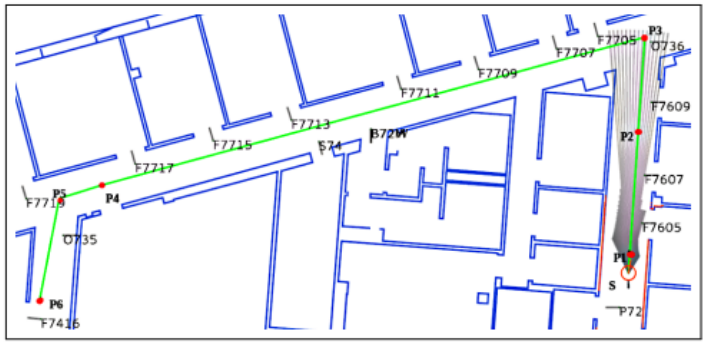
Maria Fox, Derek Long, Daniele Magazzeni: Explainable Planning. CoRR abs/1709.10256 (2017)

(Manual) Plan Comparison

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 (7)

- Robotics



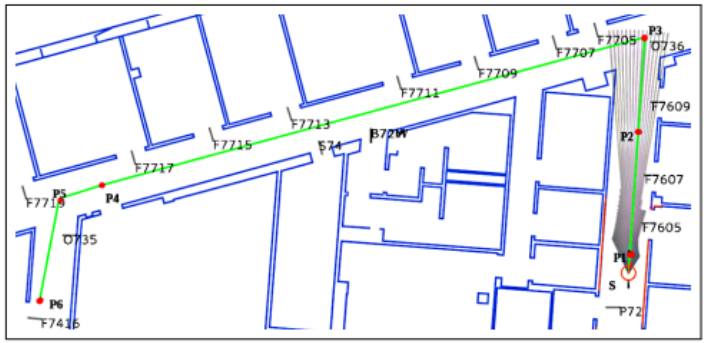
Specificity, S		Abstraction, A			
		Level 1	Level 2	Level 3	Level 4
	General Picture	Start and finish point of the complete route	Total distance and time taken for the complete route	Total distance and time taken for the complete route	Starting and ending landmark of complete route
	Summary	Start and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

Narration of Autonomous Robot Experience

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

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

Overview of explanation in different AI fields (7)



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

Narration of Autonomous Robot Experience

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

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

Robot: I have decided to turn left.

Human: Why did you do that?

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

Human: How confident are you in this decision?

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

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

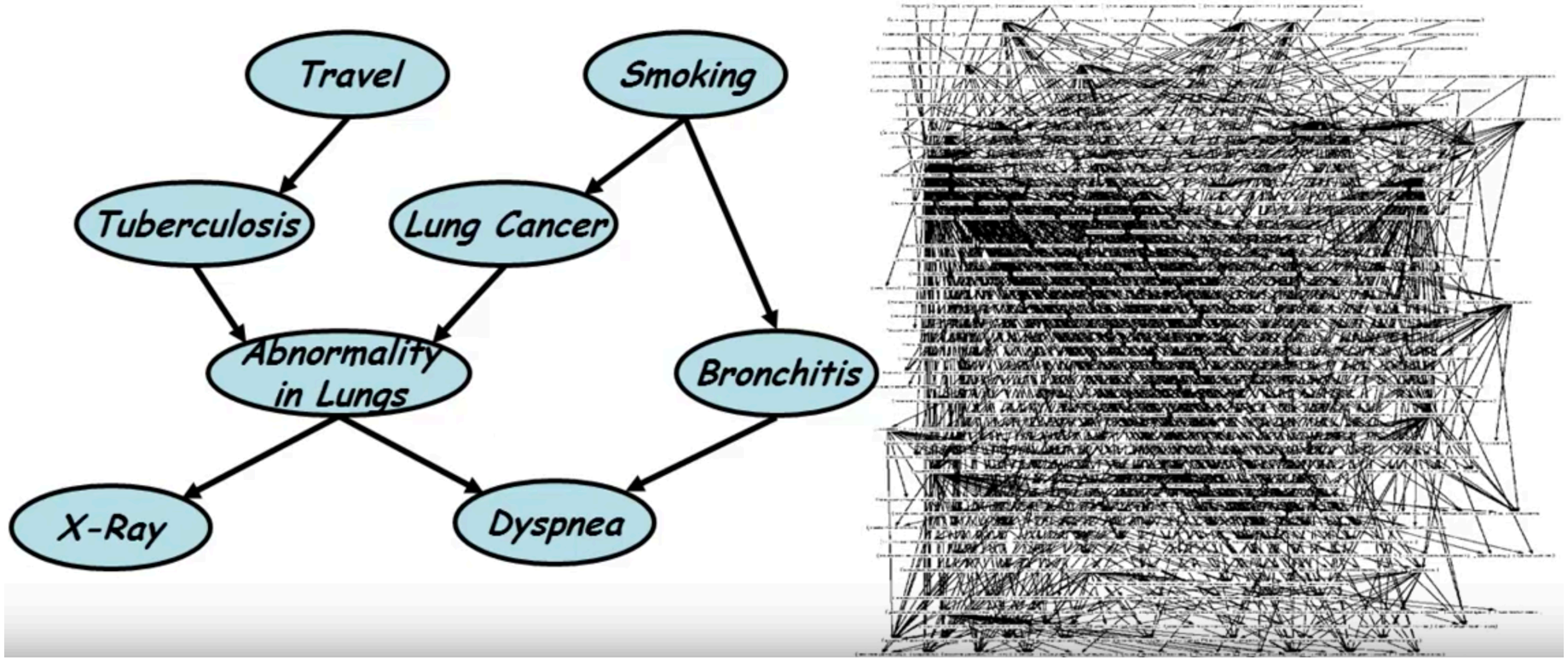
Robot: Here is the histogram of all training examples that reached this leaf. 80% of examples where this area was above 20 cm predicted the appropriate action to be "drive forward".

From Decision Tree to human-friendly information

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

Overview of explanation in different AI fields (8)

- Reasoning under Uncertainty



Probabilistic Graphical Models

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