

Statistical Machine Learning



Explainable/Interpretable AI/Machine Learning
(Version 1.0)

Hamid R. Rabiee
Spring 2023

Outline

- Introduction & Motivation
- What is Explainability/Interpretability
- Techniques for Explainability/Interpretability
- Applications of Explainability/Interpretability

Introduction & Motivation

The current state of AI (deep learning)

Game GO



Traffic Sign
Recognition



Skin cancer
detection



Lung cancer
detection



Poker



Computer games



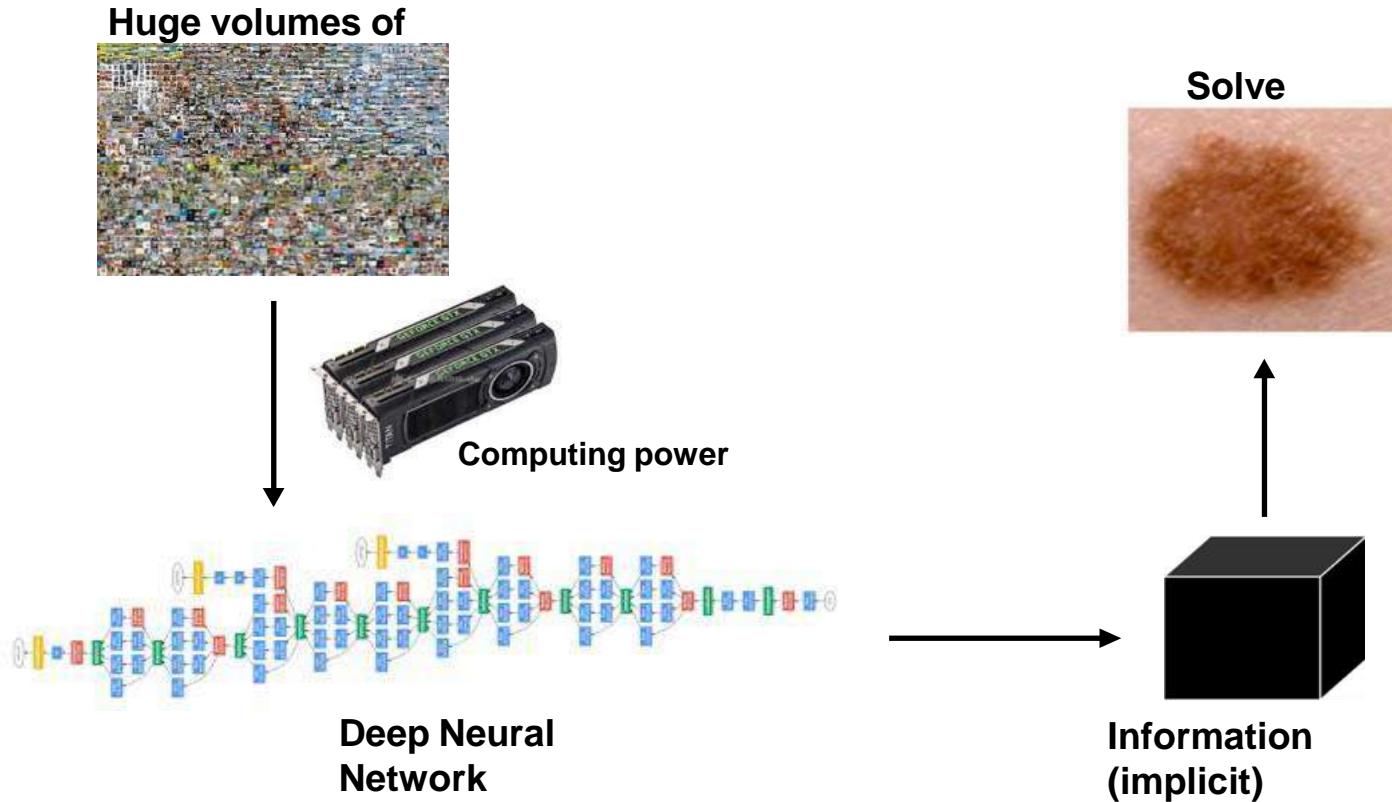
Jeopardy



OCR

Optical character recognition (OCR) is a process of conversion of images of text (e.g., photo) or from subtitle text used as a form of information bank statements, computer searched, stored more computing, machine trans in pattern recognition, ar

The current state of AI (deep learning)



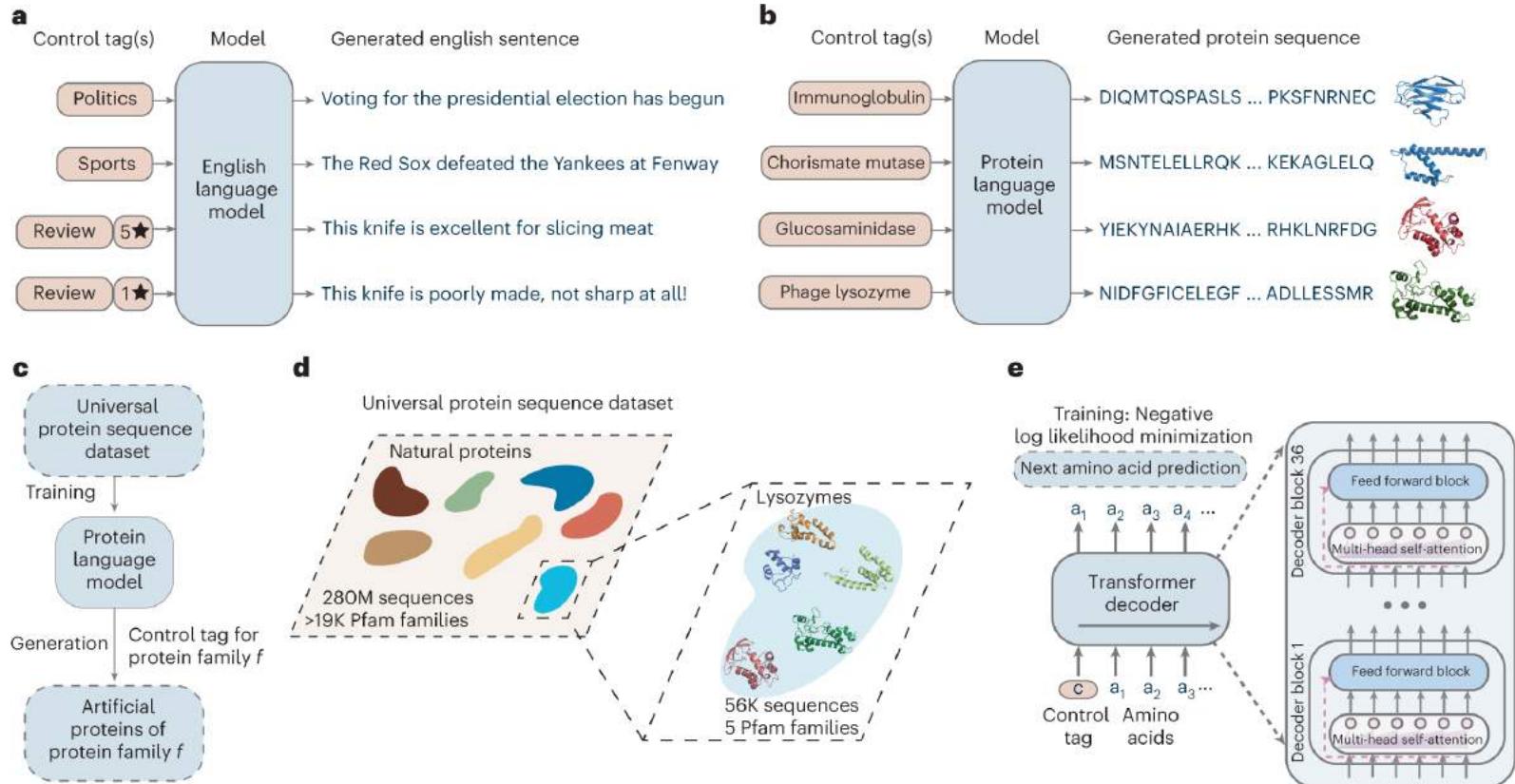
The current state of AI (LLMs)



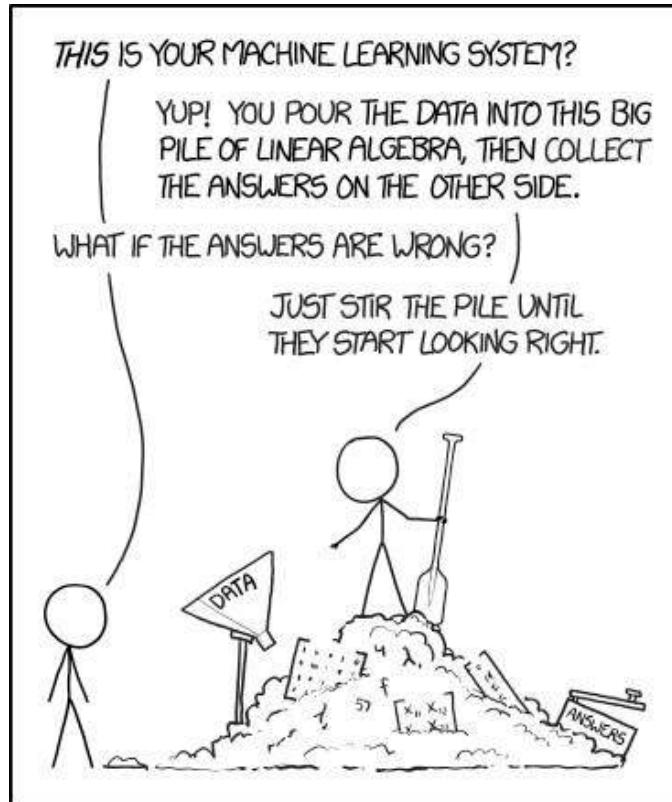
~~Google~~
ChatGPT

The word "Google" is crossed out with a large red diagonal line. Below it, the word "ChatGPT" is written in a large, bold, purple sans-serif font. Above "ChatGPT" is a purple version of the GPT swirl logo.

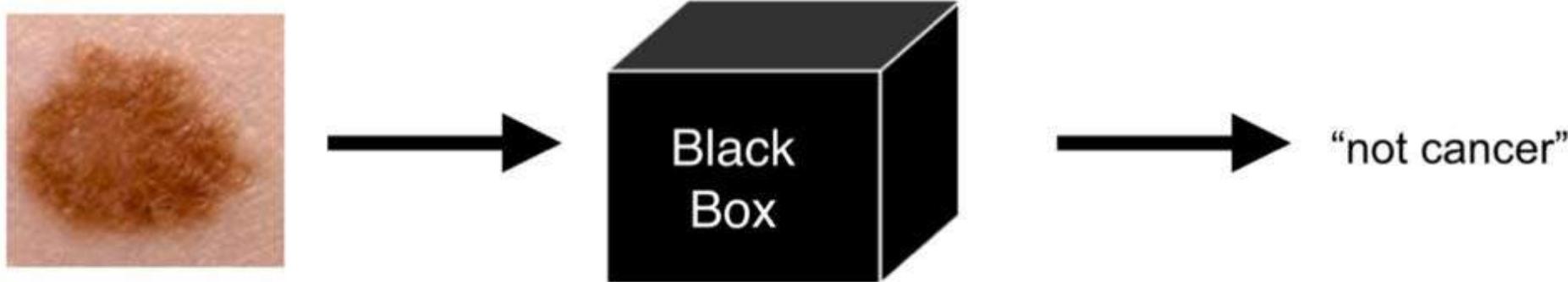
The current state of AI (LLMs)



The current state of AI (deep learning)

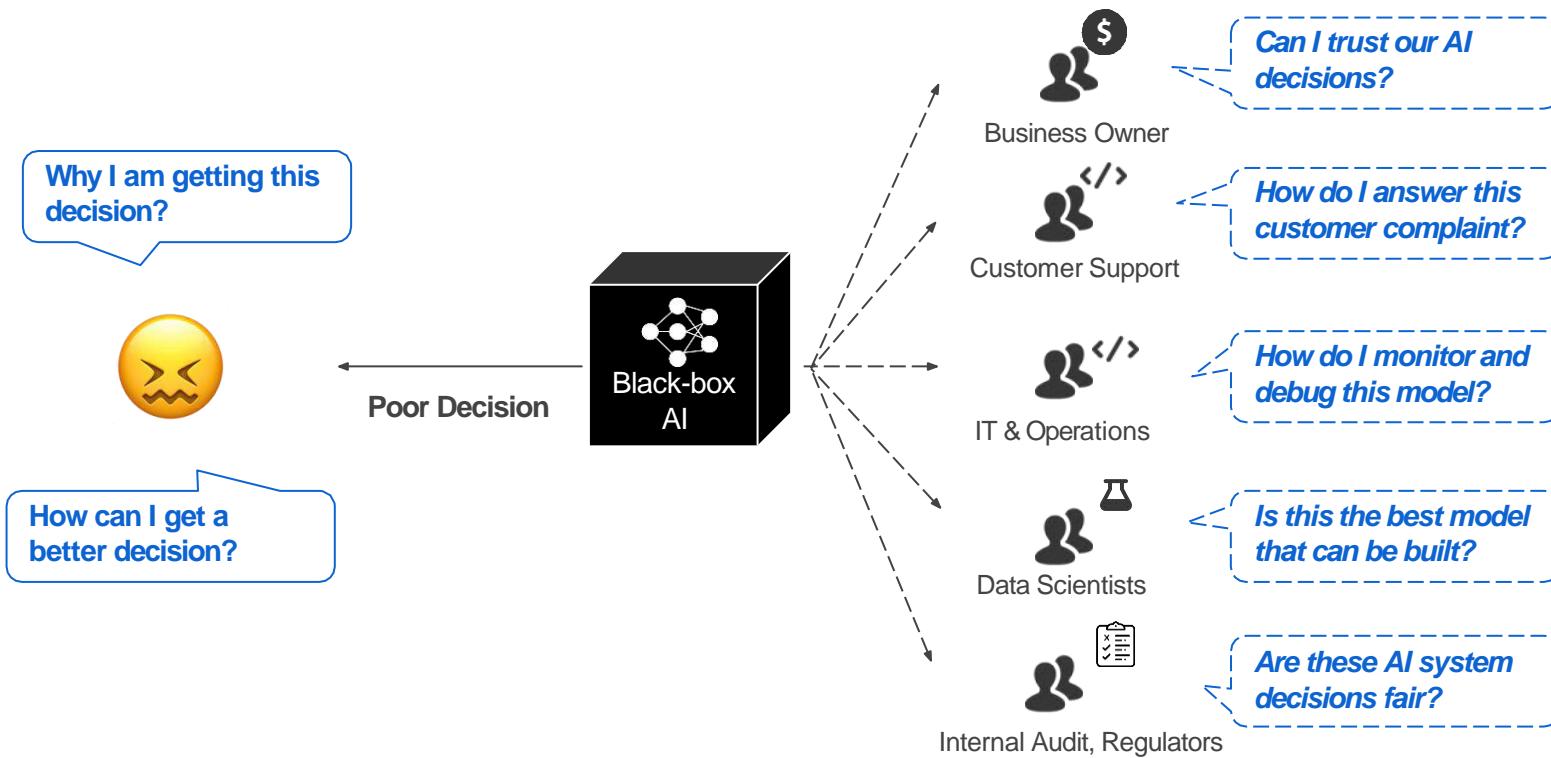


The current state of AI (deep learning)



Is minimizing the error a guarantee for the model to work well in practice?

Black-box AI creates confusion and doubt ...



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



Safety and well being

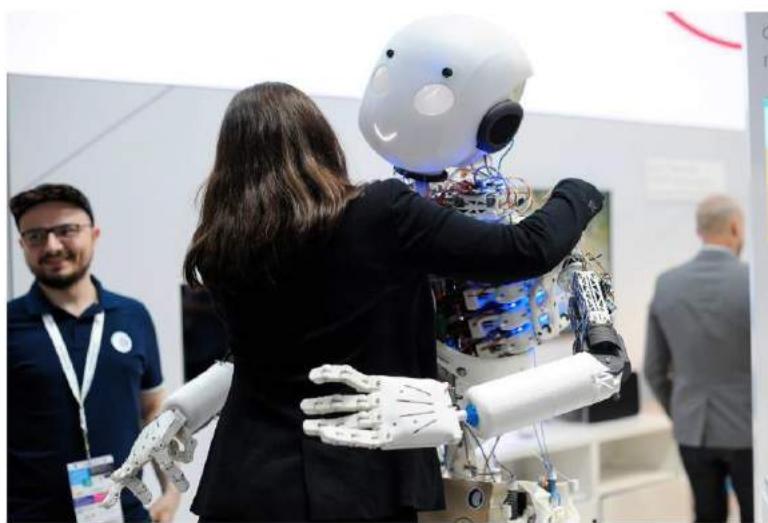
Tesla hit parked police car 'while using Autopilot'

© 30 May 2018

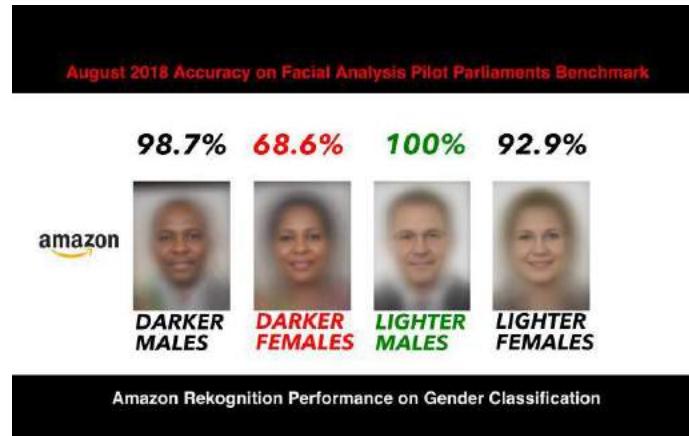
f t tw em Share



*Warnings of a Dark Side
to A.I. in Health Care*



Bias in algorithms



<https://medium.com/@Joy.Buolamwini/response-racial-and-gender-bias-in-amazon-rekognition-commercial-ai-system-for-analyzing-faces-a289222eeecd>

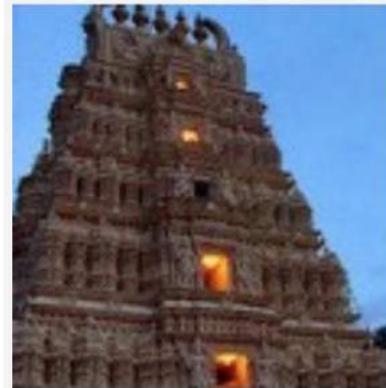
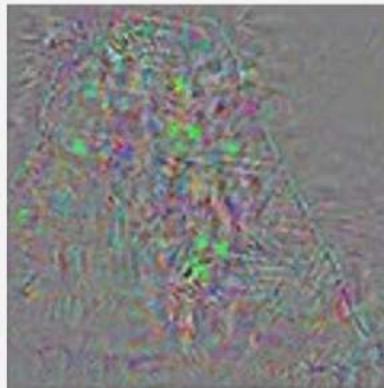
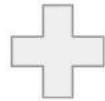
Machine Learning can amplify bias.



- Data set: 67% of people cooking are women
- Algorithm predicts: 84% of people cooking are women

<https://www.infoq.com/presentations/unconscious-bias-machine-learning/>

Adversarial Examples



Original image

Temple (97%)

Perturbations

Adversarial example

Ostrich (98%)

Legal Issues – GDPR (General Data Protection Regulation)



Pedro Domingos
@pmddomingos

Follow

Starting May 25, the European Union will require algorithms to explain their output, making deep learning illegal.

7:59 PM - 28 Jan 2018

188 Retweets 312 Likes



41

188

312



GDPR Concerns Around Lack of Explainability in AI

“

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

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

”

- European Commission



Andrus Ansip
@Ansip_EU

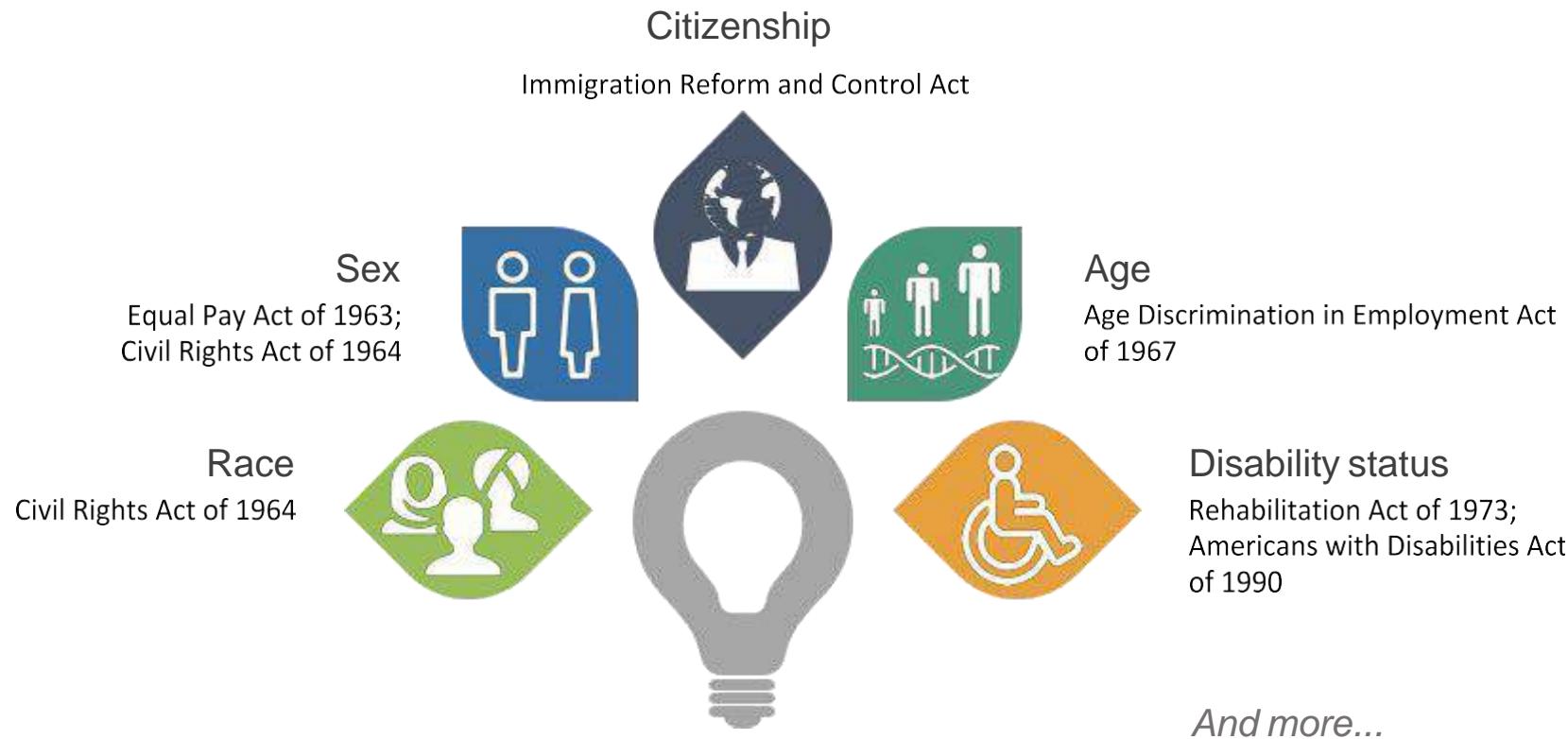
You have the right to be informed about an automated decision and ask for a human being to review it, for example if your online credit application is refused.
#EUdataP #GDPR #AI #digitalrights
#EUandMe europa.eu/!nN77Dd



8:30 AM - 7 Sep 2018

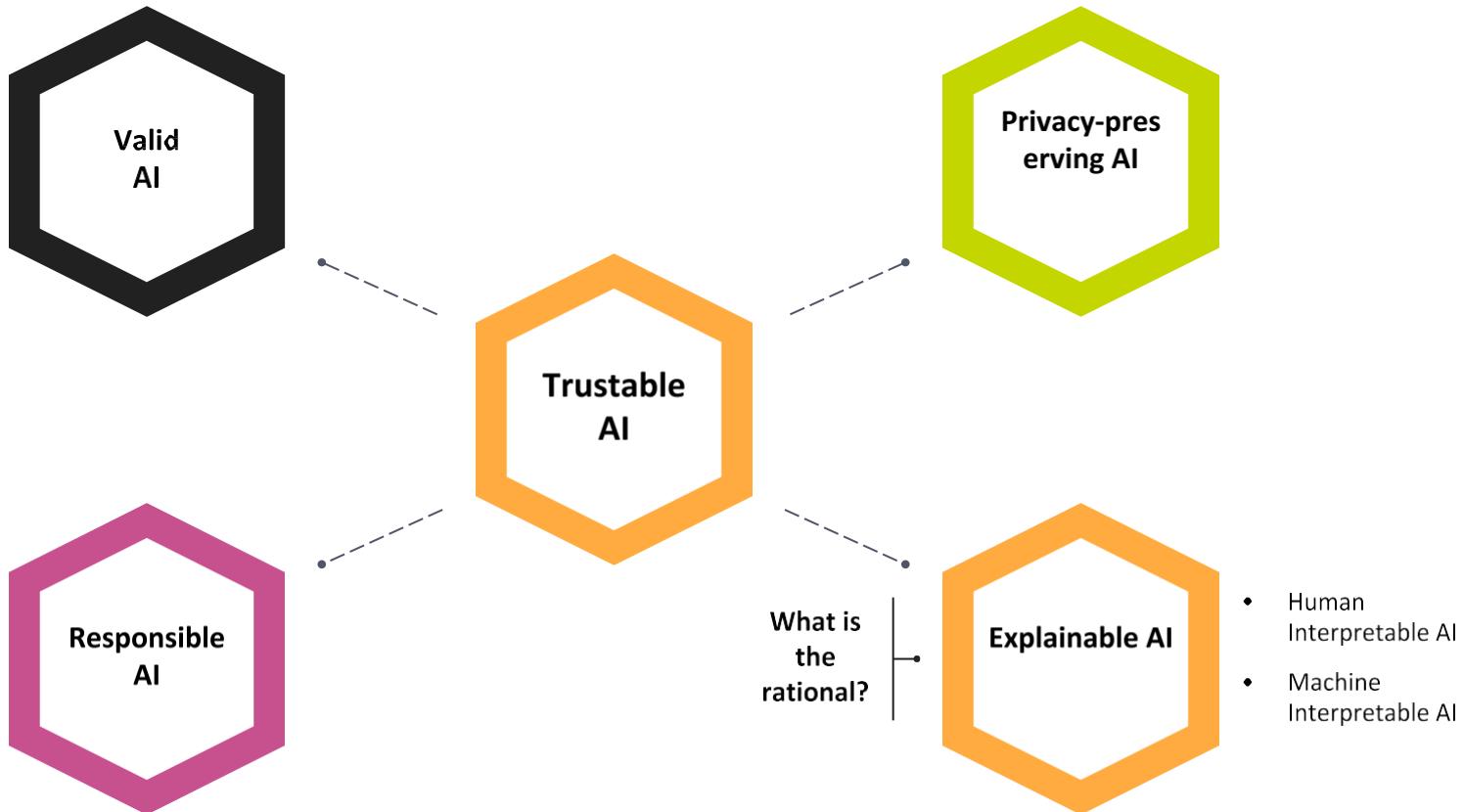
VP, European Commission

Why Explainability: Laws against Discrimination



What is Explainability/Interpretability

AI/ML Adoption: Requirements



AI/ML Adoption: Requirements

Trustworthy AI is valid and reliable, safe, fair, and bias is managed, secure and resilient, accountable and transparent, explainable and interpretable, and privacy-enhanced.

Valid AI refers to the accuracy of a measure; whether the results really do represent what they are supposed to measure.

Responsible AI is the practice of designing, developing, and deploying AI with good intention.

Privacy-preserving refers to merging techniques that help preserve the user's privacy. The building blocks of privacy-preserving machine learning are federated learning, homomorphic Encryption, and differential privacy.

Interpretability has to do with how accurate a machine learning model can associate a cause to an effect. **Explainability** has to do with the ability of the parameters, often hidden in Deep Nets, to justify the results.

For AI/ML methods, the terms **interpretability and explainability** are commonly interchangeable.

There is no standard definition!

- Most agree it is something different from performance.
- Ability to explain or to present a model in understandable terms to humans (Doshi-Velez 2017)
- Cynical view – It is what makes you feel good about the model.
- It really depends on target audience.

XAI Definitions - Explanation vs. Interpretation

explanation | ɛksplə'neɪʃ(ə)n |

noun

Oxford Dictionary
of English

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

interpret | ɪn'təprɪt |

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

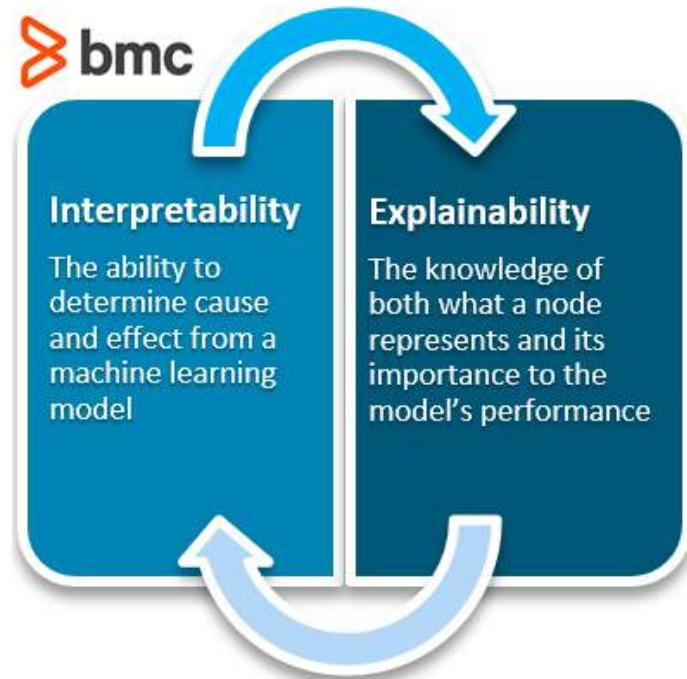
1 explain the meaning of (information or actions): *the evidence is difficult to interpret.*

XAI Definitions - Explanation vs. Interpretation

- Explainable AI (XAI), also known as Interpretable AI, or Explainable Machine Learning (XML), is artificial intelligence in which humans can understand the reasoning behind decisions or predictions made by the AI.
- Explainable AI models “summarize the reasons [...] for [their] behavior [...] or produce insights about the causes of their decisions,”
- Whereas Interpretable AI refers to AI systems which “describe the internals of a system in a way which is understandable to humans.”

XAI Definitions - Explanation vs. Interpretation

- Explainable AI tells you **why** it made the decision it did, but **not how** it arrived at that decision.
- Interpretable AI tells you **how** it made the decision, but **not why** the criteria it used is sensible.
- We can of course imagine systems that are both Explainable and Interpretable.



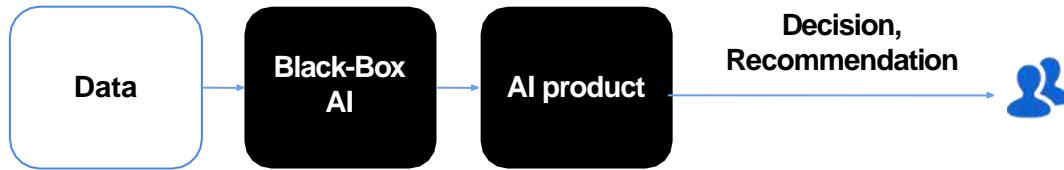
In general, it seems like there are few fundamental problems:

- We don't trust the models
- We don't know what happens in extreme cases
- Mistakes can be expensive / harmful
- Does the model makes similar mistakes as humans ?
- How to change model when things go wrong ?

Interpretability is one way we try to deal with these problems

What is Explainable AI?

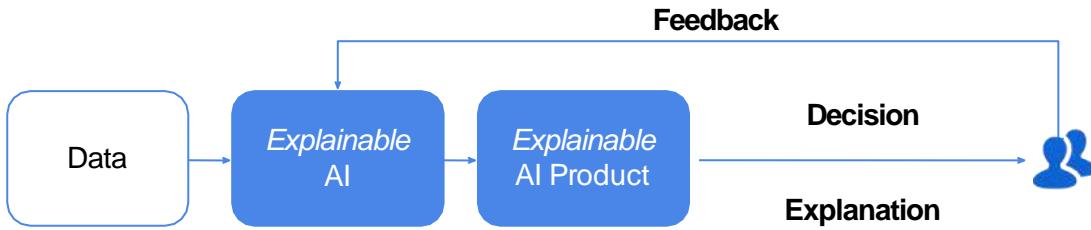
Black Box AI



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

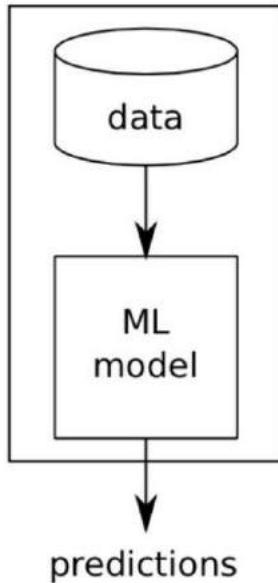


Clear & Transparent Predictions

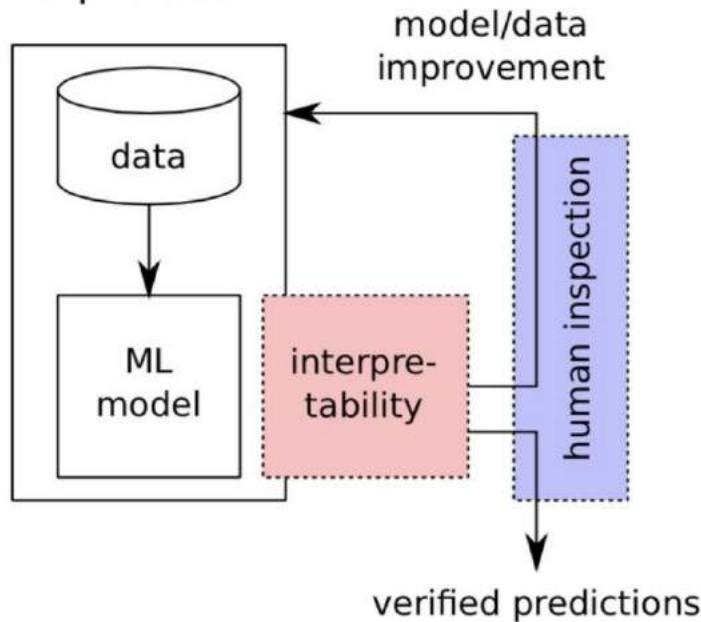
- I understand why
- I understand why not
- I know why you succeed or fail
- I understand, so I trust you

Why Explainability: Improve ML Model

Standard ML



Interpretable ML



Generalization error

Generalization error + human experience

And more ...

- **Interactive feedback** - can model learn from human actions in online setting ? (Can you tell a model to not repeat a specific mistake ?)
- **Recourse** – Can a model tell us what actions we can take to change its output ? (For example, what can you do to improve your credit score?)

What does interpretation looks like?

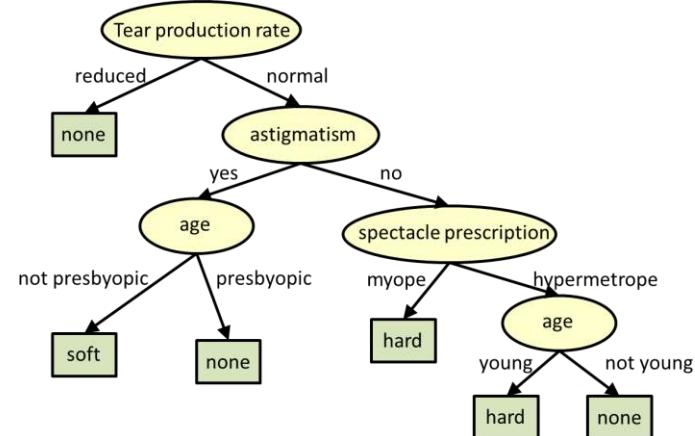
- In pre-deep learning models, some models are considered “interpretable”

Dependent Variable →

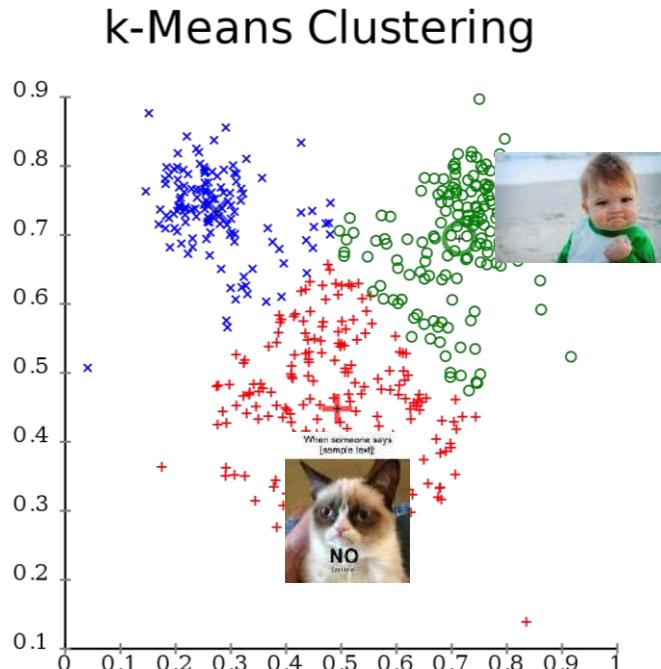
$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

Population Y intercept Population Slope Coefficient Independent Variable Random Error term

Linear component Random Error component



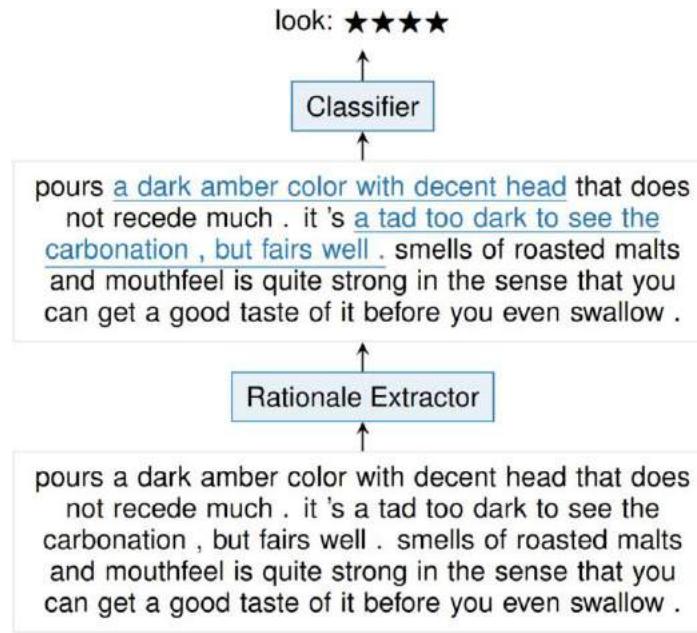
What does interpretation looks like?



By Chire - Own work, Public Domain,
<https://commons.wikimedia.org/w/index.php?curid=11765684>

What does interpretation look like?

- Bake it into the model



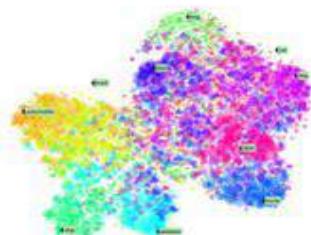
[Bastings et al 2019]

Some properties of Interpretations

- **Faithfulness** - how to provide explanations that accurately represent the true reasoning behind the model's final decision.
- **Plausibility** – Is the explanation correct or something we can believe is true, given our current knowledge of the problem?
- **Understandable** – Can I put it in terms that end user without in-depth knowledge of the system can understand?
- **Stability** – Does similar instances have similar interpretations ?

Dimensions of Interpretability

Different dimensions
of “interpretability”

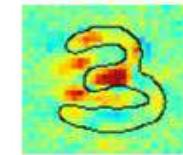


data

*“Which dimensions of the data
are most relevant for the task.”*

prediction

*“Explain why a certain pattern x has
been classified in a certain way $f(x)$.”*



model

*“What would a pattern belonging
to a certain category typically look
like according to the model.”*



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

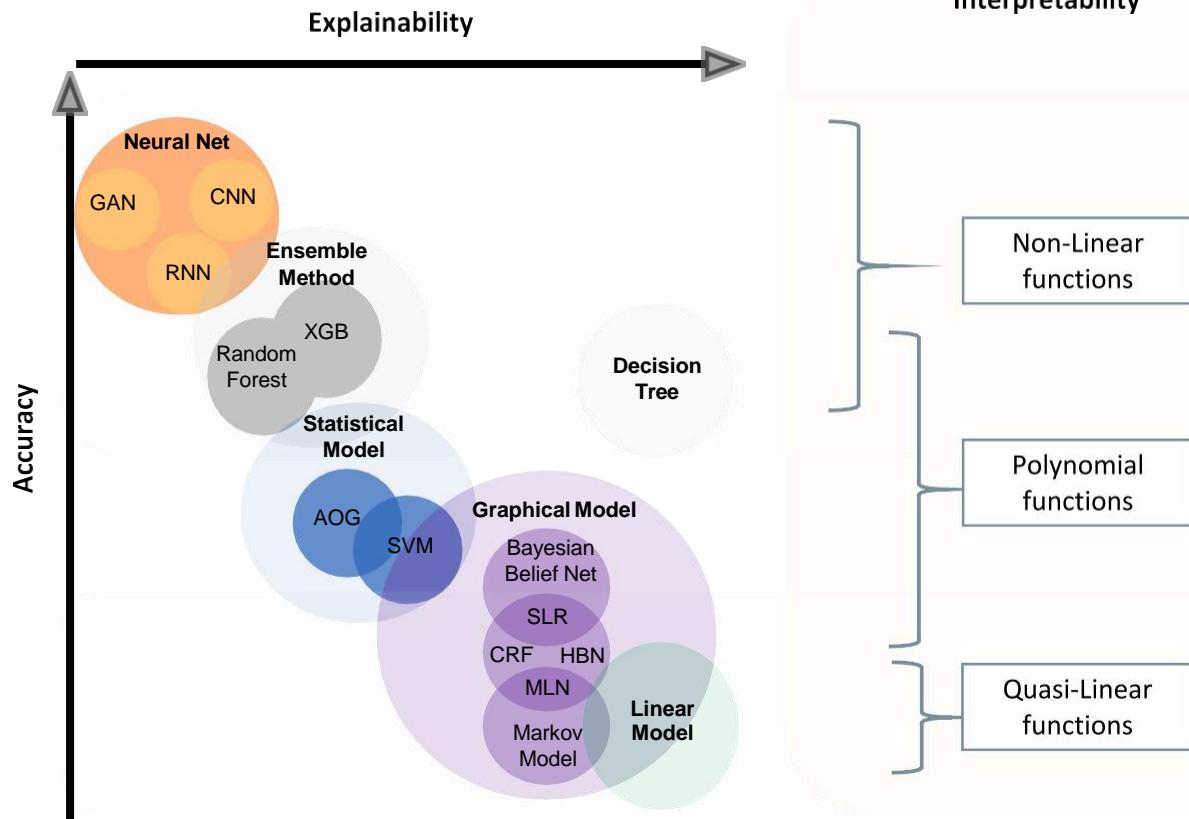


Text

How to Explain? Accuracy vs. Explainability

Learning

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



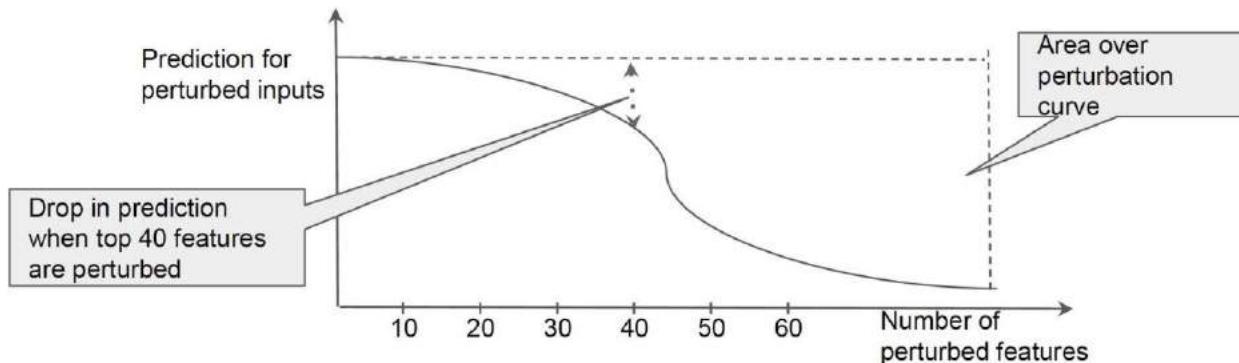
Evaluating Interpretability [Doshi-Velez 2017]

- **Application level evaluation**: Put the model in practice and have the end users interact with explanations to see if they are useful.
- **Human evaluation**: Set up a Mechanical Turk task and ask non-experts to judge the explanations.
- **Functional evaluation**: Design metrics that directly test properties of your explanation.

Evaluating Interpretability: 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



Evaluating Interpretability: 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?

Evaluating Interpretability: 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?



Techniques for Explainability/Interpretability

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)

train interpretable
model

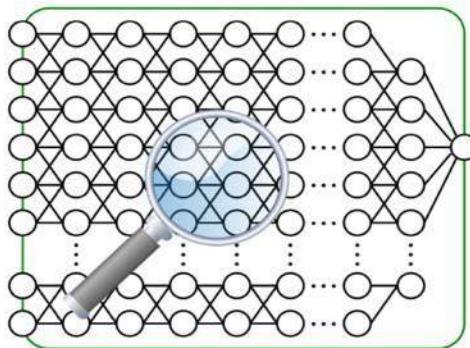
*suboptimal or biased due to
assumptions (linearity, sparsity ...)*

vs.

train best
model → interpret it

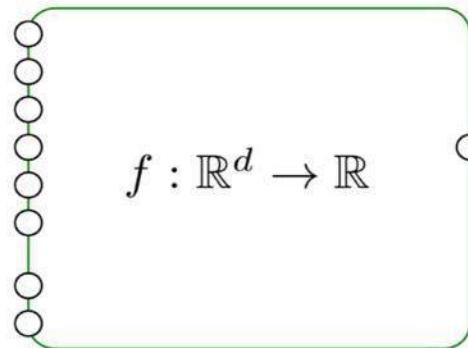
Techniques of Interpretation

**mechanistic
understanding**



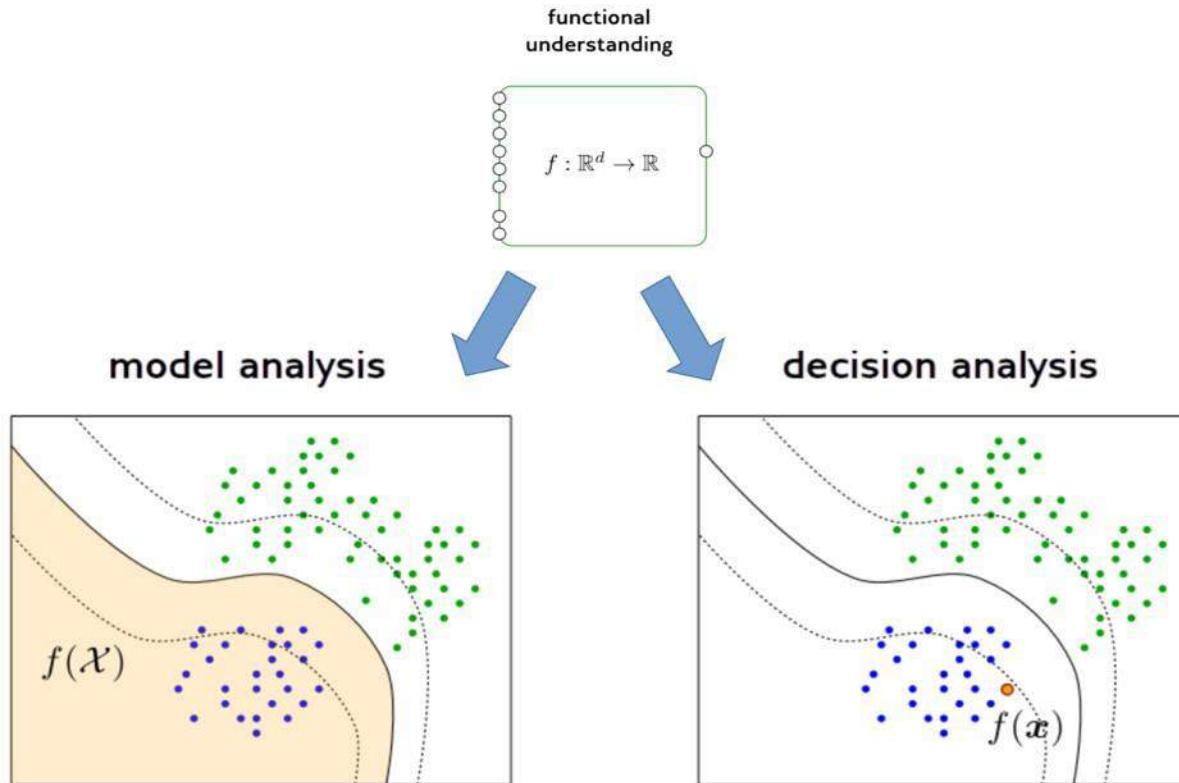
Understanding what mechanism the network uses to solve a problem or implement a function.

**functional
understanding**



Understanding how the network relates the input to the output variables.

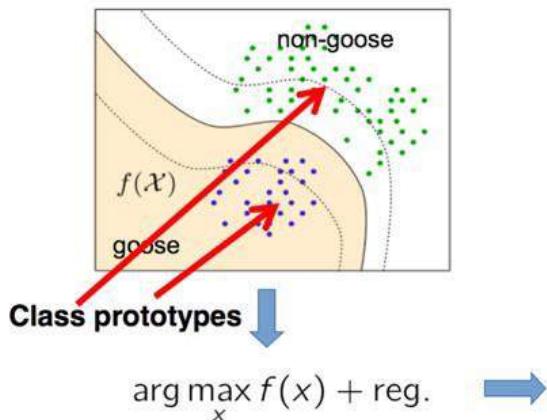
Techniques of Interpretation



Interpreting the Model

Approach 1: Class Prototypes

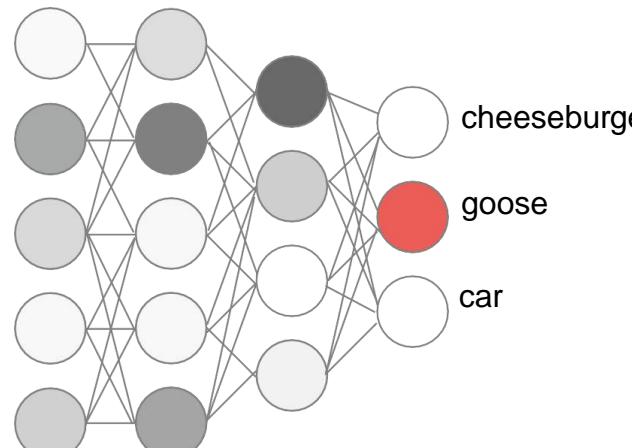
"How does a goose typically look like according to the neural network?"



Interpreting the Model

Activation Maximization

- find prototypical example of a category
- find pattern maximizing activity of a neuron

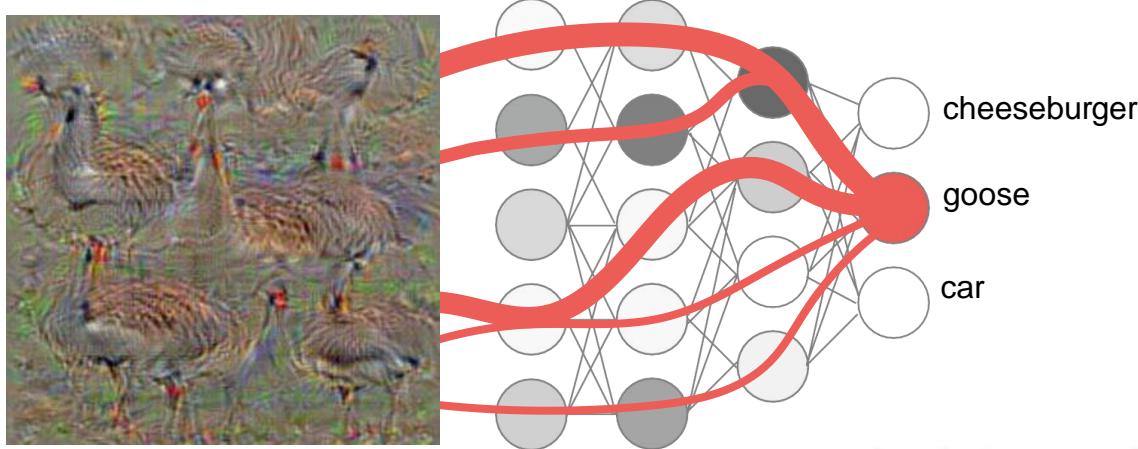


$$\max_{x \in \mathcal{X}} p_\theta(\omega_c | x) + \lambda \Omega(x)$$

Interpreting the Model

Activation Maximization

- find prototypical example of a category
- find pattern maximizing activity of a neuron



simple regularizer
(Simonyan et al.
2013)

$$\max_{x \in \mathcal{X}} p_\theta(\omega_c | x) + \lambda \Omega(x)$$

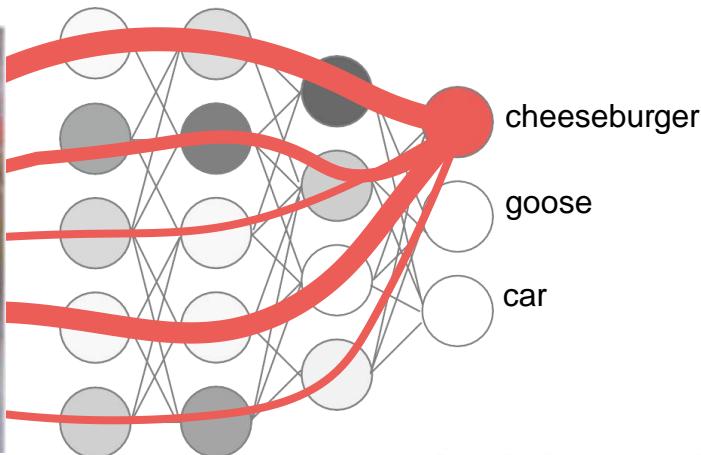
Interpreting the Model

Activation Maximization

- find prototypical example of a category
- find pattern maximizing activity of a neuron



complex regularizer
(Nguyen et al. 2016)

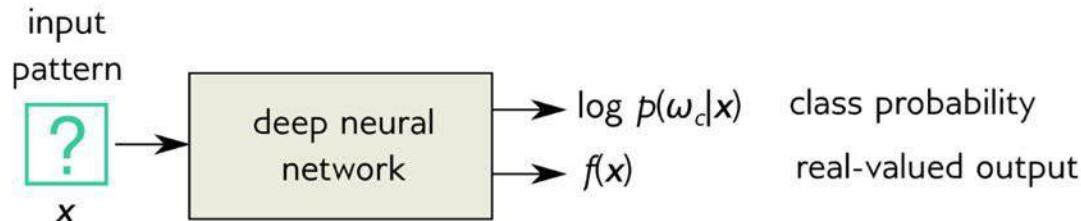


$$\max_{x \in \mathcal{X}} p_\theta(\omega_c | x) + \lambda \Omega(x)$$

Interpreting the Model

Activation Maximization

Let us interpret a concept predicted by a deep neural net (e.g. a class, or a real-valued quantity):



Examples:

- ▶ Creating a class prototype: $\max_{x \in \mathcal{X}} \log p(\omega_c|x)$.
- ▶ Synthesizing an extreme case: $\max_{x \in \mathcal{X}} f(x)$.

Interpreting the Model



Images from **Simonyan et al. 2013** “Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps”

Observations:

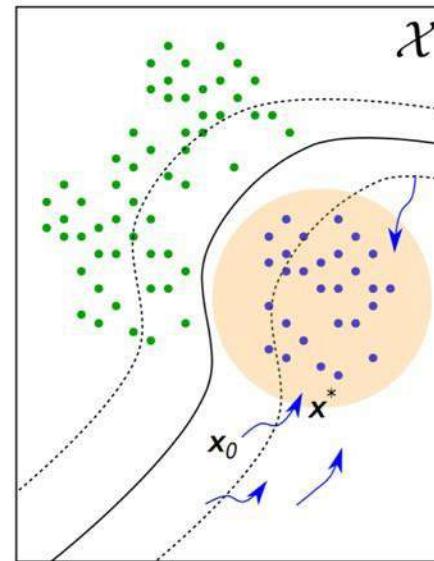
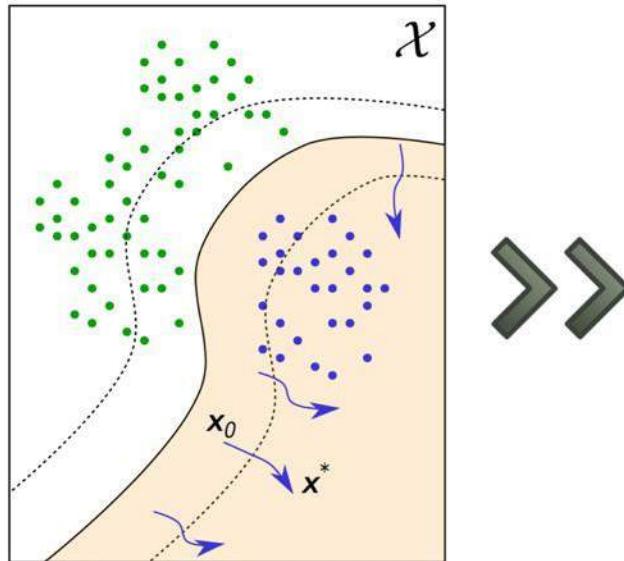
- ▶ AM builds typical patterns for these classes (e.g. beaks, legs).
- ▶ Unrelated background objects are not present in the image.

Interpreting the Model: Enhancing Activation Maximization

Find the input pattern that maximizes class probability.

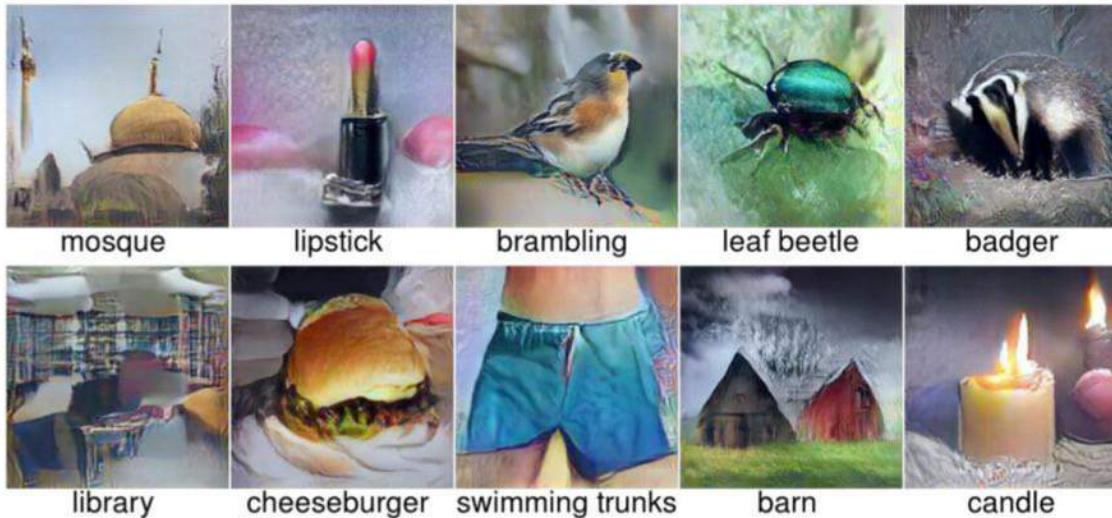


Find the most likely input pattern for a given class.



Interpreting the Model: Enhancing Activation Maximization

Images from Nguyen et al. 2016. "Synthesizing the preferred inputs for neurons in neural networks via deep generator networks"



Observation: Connecting AM to the data distribution leads to more realistic and more interpretable images.

Limitations of Global Interpretations

Question: Below are some images of motorbikes. What would be the best prototype to interpret the class “motorbike”?

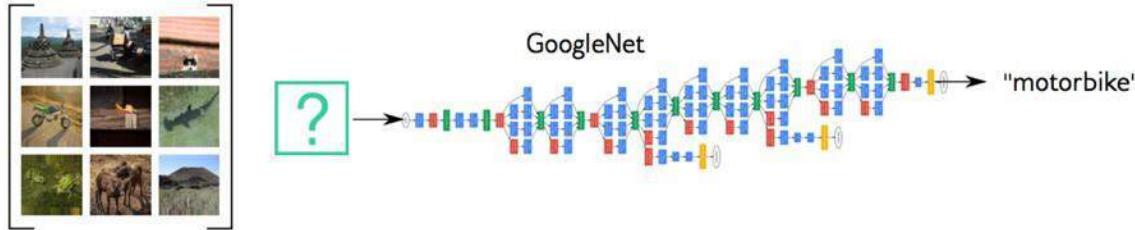


Observations:

- ▶ Summarizing a concept or category like “motorbike” into a single image can be difficult (e.g. different views or colors).
- ▶ A good interpretation would grow as large as the diversity of the concept to interpret.

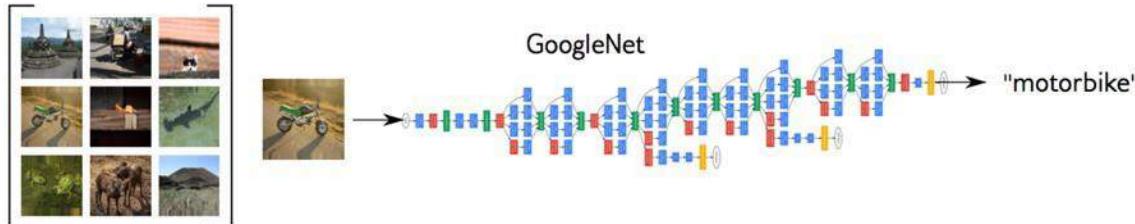
Need for Individual Explanations

Finding a prototype:



Question: How does a “motorbike” typically look like?

Individual explanation:



Question: Why is *this* example classified as a motorbike?

Need for Individual Explanations

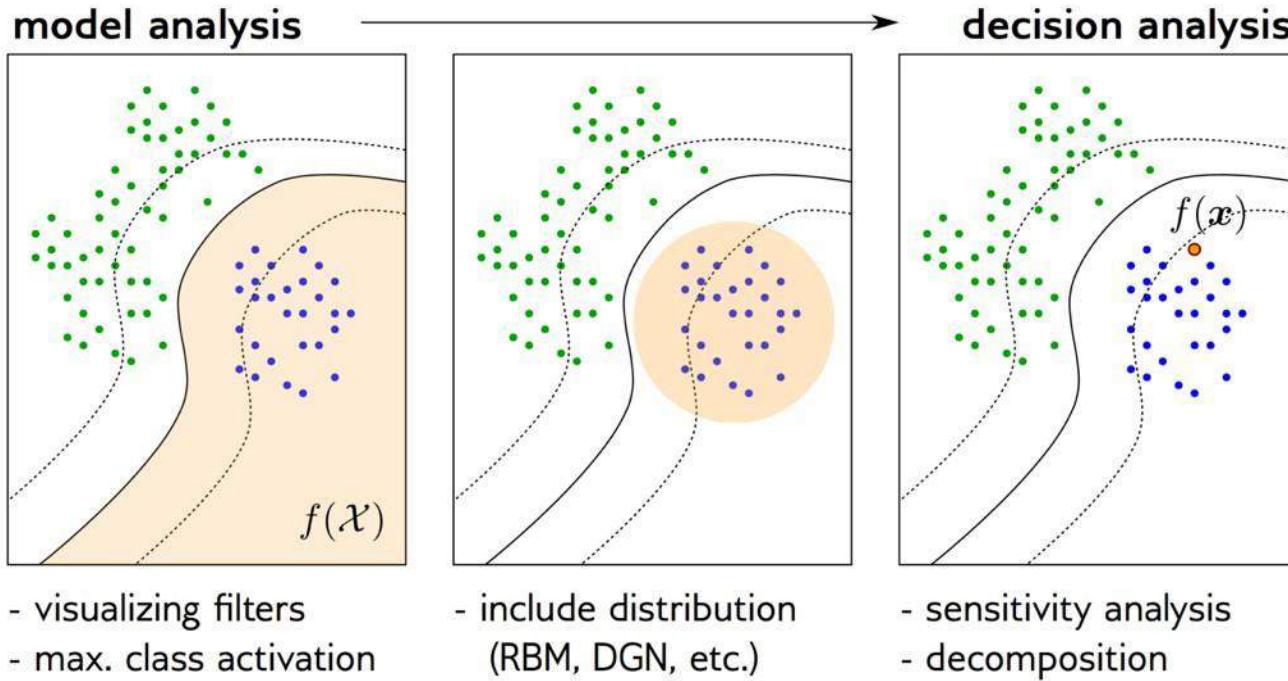
Personalized medicine: Extracting the relevant information about a medical condition for a *given* patient at a *given* time.

Each case is unique and
needs its own explanation.

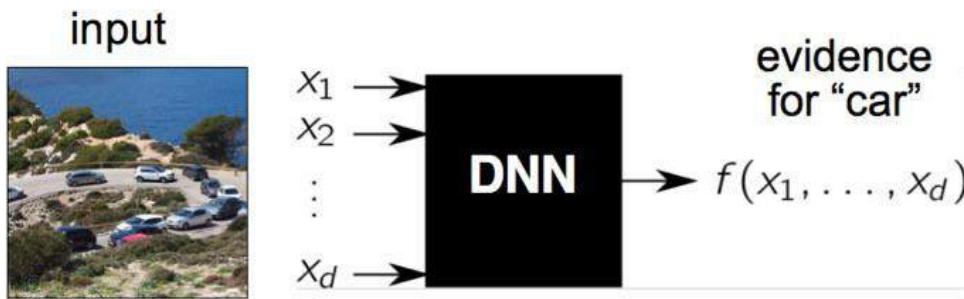
Population view: Which symptoms are most common for the disease

Both aspects can be important depending on who you are
(FDA, doctor, patient).

Making Deep Neural Nets Transparent



Decision Analysis: Sensitivity Analysis



Sensitivity analysis: The relevance of input feature i is given by the squared partial derivative:

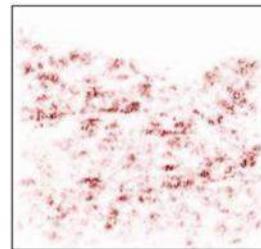
$$R_i = \left(\frac{\partial f}{\partial x_i} \right)^2$$

Decision Analysis: Sensitivity Analysis

Sensitivity analysis:



$$R_i = \left(\frac{\partial f}{\partial x_i} \right)^2$$



Problem: sensitivity analysis does not highlight cars

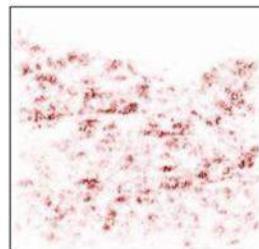
highlights parts, which (when changed) increase or decrease the prediction for “car”.

Decision Analysis: Sensitivity Analysis

Sensitivity analysis:



$$R_i = \left(\frac{\partial f}{\partial x_i} \right)^2$$



Problem: sensitivity analysis does not highlight cars

highlights parts, which (when changed) increase or decrease the prediction for “car”.

Observation:

$$\sum_{i=1}^d \left(\frac{\partial f}{\partial x_i} \right)^2 = \|\nabla_x f\|^2$$

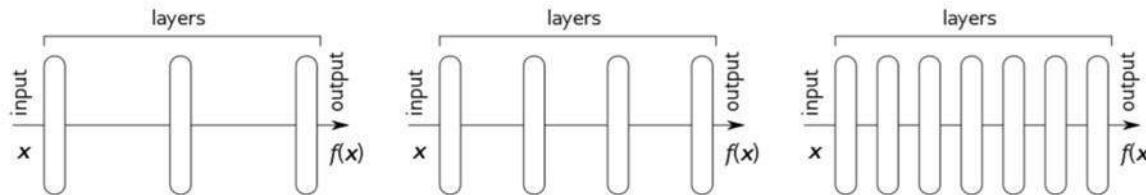
Sensitivity analysis explains a *variation* of the function, not the function value itself.

Decision Analysis: Sensitivity Analysis

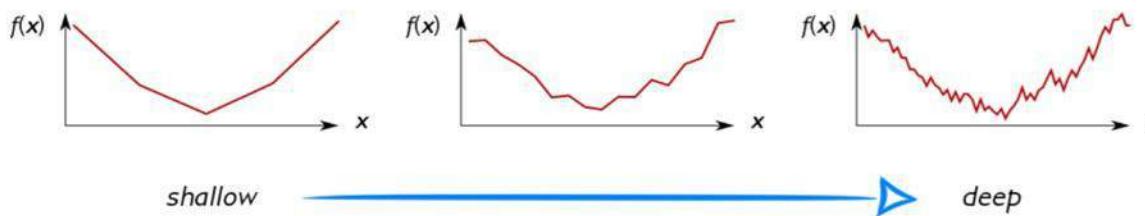
Shattered Gradient Problem

Input gradient (on which sensitivity analysis is based), becomes increasingly highly varying and unreliable with neural network depth.

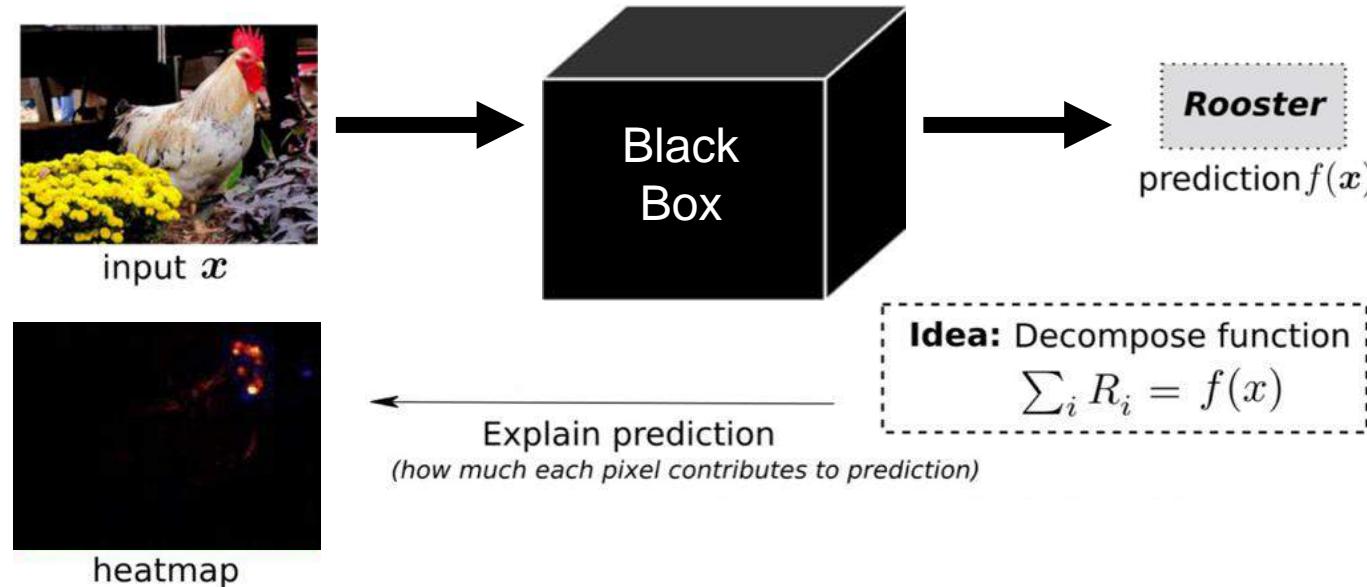
Structure's view



Function's view (cartoon)



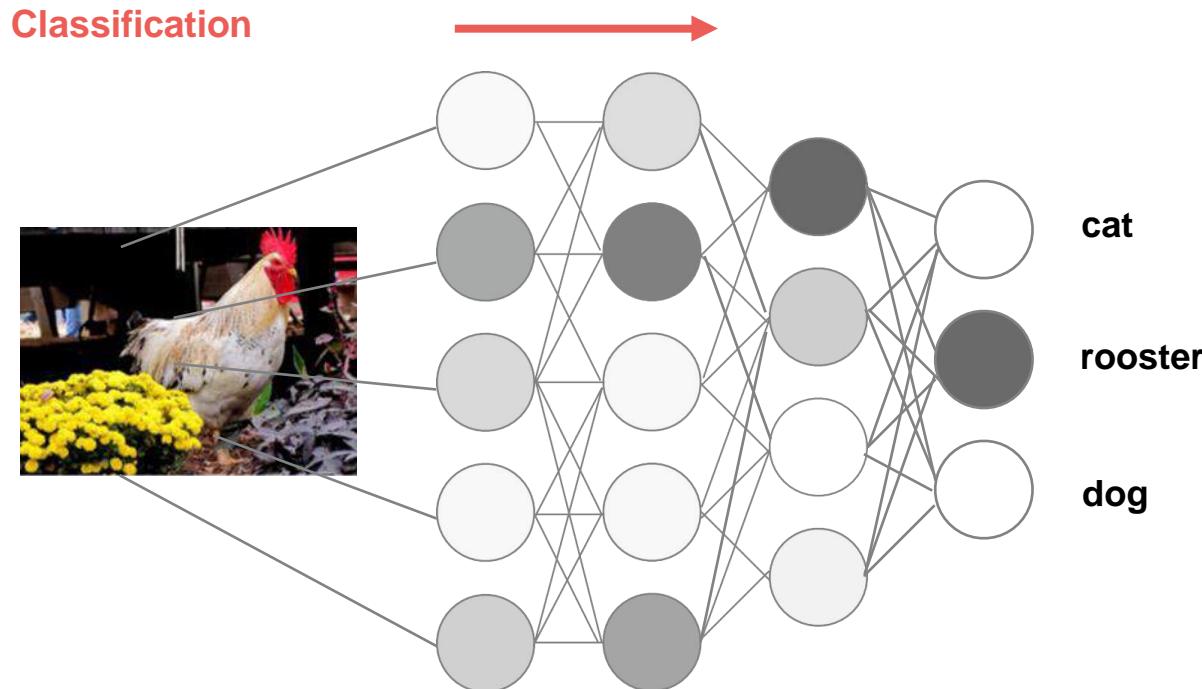
Decision Analysis: LRP (Layer-wise Relevance Propagation)



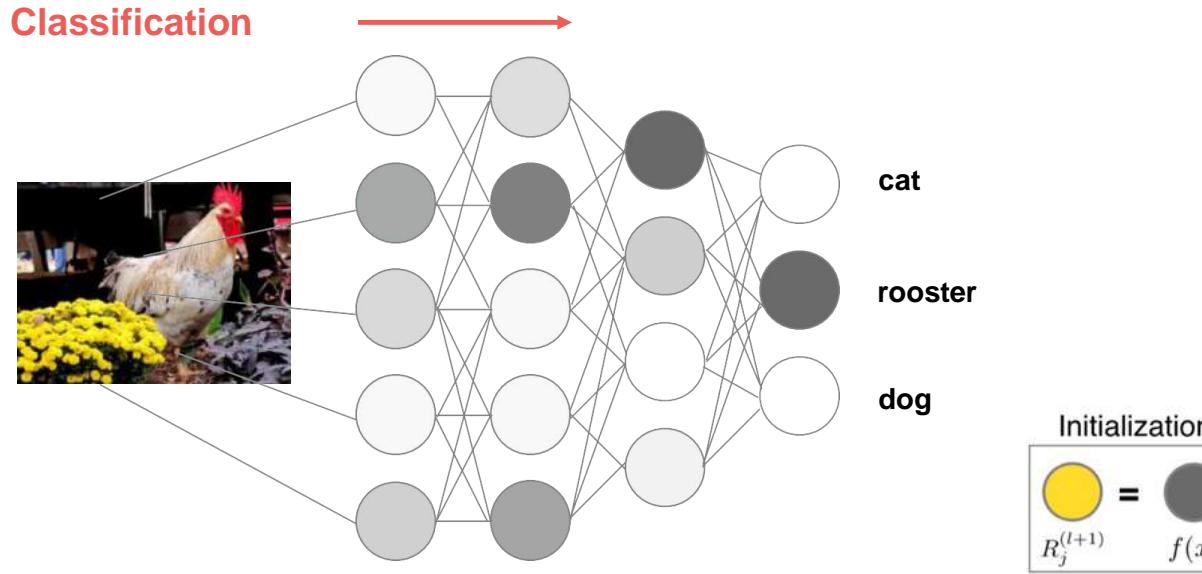
Layer-wise Relevance Propagation
(LRP) (Bach et al., PLOS ONE, 2015)

Explain prediction itself
(not the change)

Decision Analysis: LRP (Layer-wise Relevance Propagation)



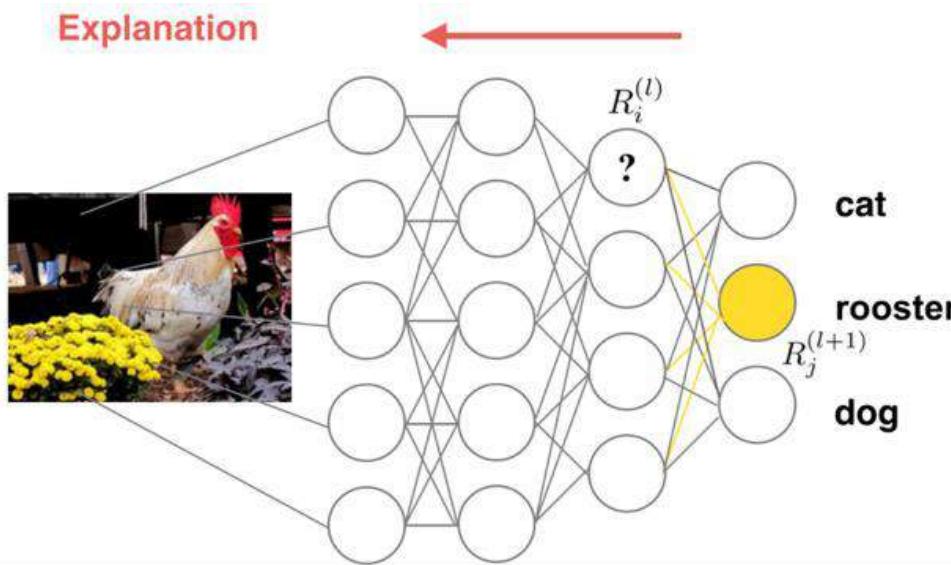
Decision Analysis: LRP (Layer-wise Relevance Propagation)



What makes this image
a “rooster image”?

Idea: Redistribute the evidence for
class rooster back to image space.

Decision Analysis: LRP (Layer-wise Relevance Propagation)

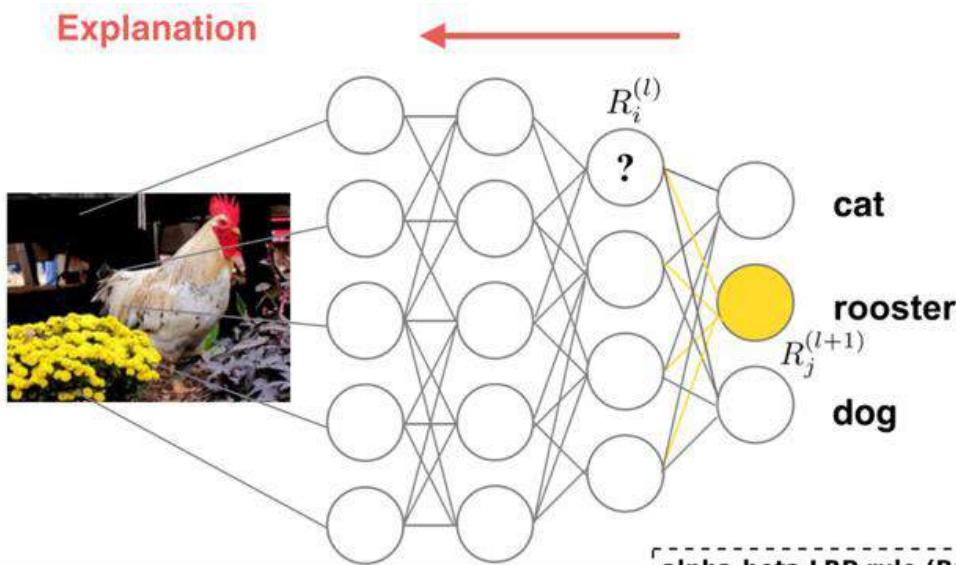


Simple LRP rule (Bach et al. 2015)

$$R_i^{(l)} = \sum_j \frac{x_i \cdot w_{ij}}{\sum_{i'} x_{i'} \cdot w_{i'j}} R_j^{(l+1)}$$

Every neuron gets its "share"
of the redistributed relevance

Decision Analysis: LRP (Layer-wise Relevance Propagation)



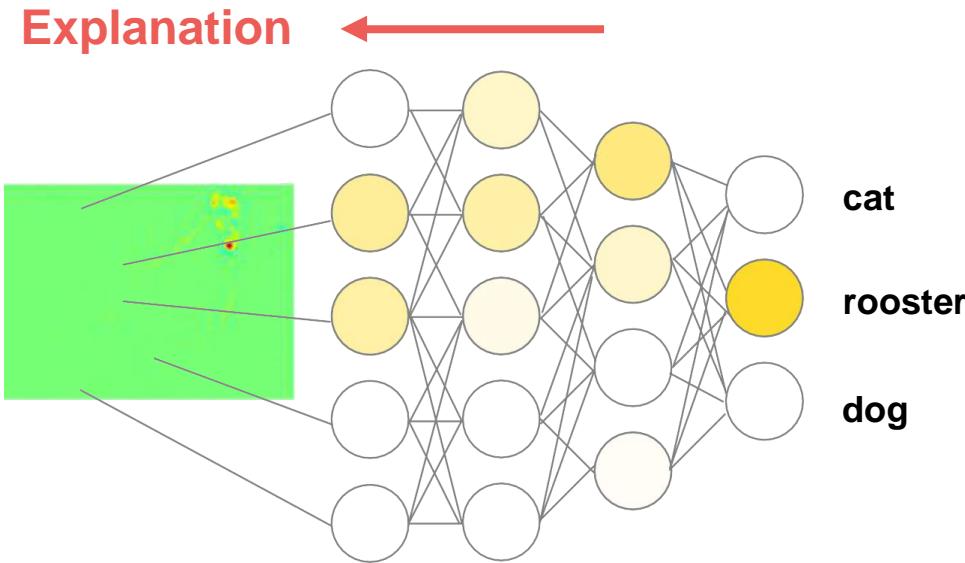
Theoretical interpretation Deep Taylor Decomposition (Montavon et al., 2017)
(no gradient shattering)

alpha-beta LRP rule (Bach et al. 2015)

$$R_i^{(l)} = \sum_j (\alpha \cdot \frac{(x_i \cdot w_{ij})^+}{\sum_{i'} (x_{i'} \cdot w_{i'j})^+} + \beta \cdot \frac{(x_i \cdot w_{ij})^-}{\sum_{i'} (x_{i'} \cdot w_{i'j})^-}) R_j^{(l+1)}$$

where $\alpha + \beta = 1$

Decision Analysis: LRP (Layer-wise Relevance Propagation)

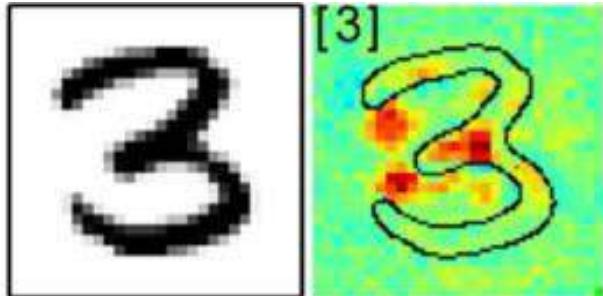


Layer-wise relevance conservation

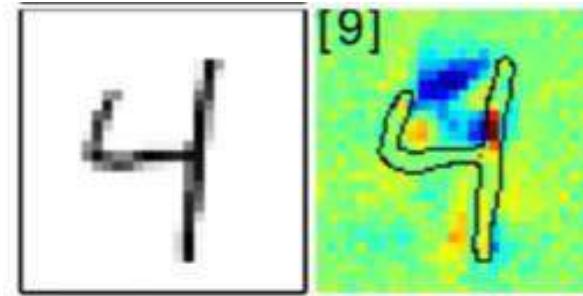
$$\sum_i R_i = \dots = \sum_i R_i^{(l)} = \sum_j R_j^{(l+1)} = \dots = f(x)$$

Decision Analysis: LRP (Layer-wise Relevance Propagation)

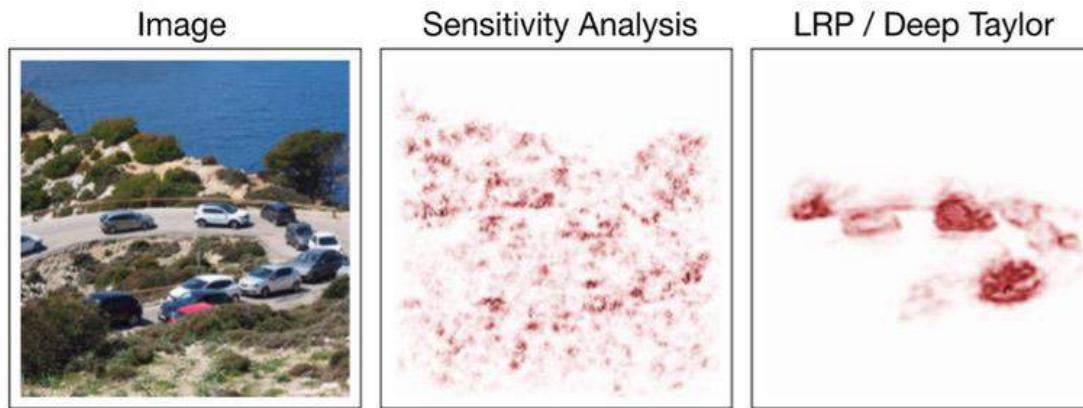
Heatmap of prediction “3”



Heatmap of prediction “9”



Decision Analysis: LRP (Layer-wise Relevance Propagation)



Explains what influences prediction “cars”.

Slope decomposition

$$\sum_i R_i = \|\nabla_x f\|^2$$

Explains prediction “cars” as is.

Value decomposition

$$\sum_i R_i = f(\mathbf{x})$$

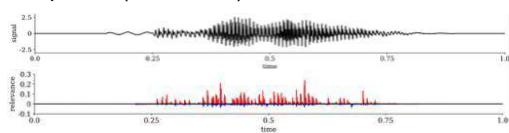
More information
(Montavon et al., 2017 & 2018)

Summary LRP

General Images (Bach' 15, Lapuschkin'16)



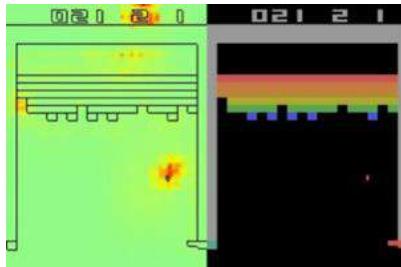
Speech (Becker'18)



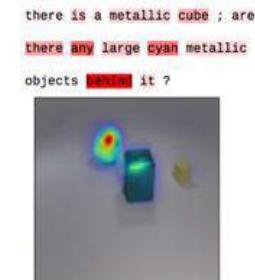
Text Analysis (Arras'16 &17)

do n't waste your money
neither funny nor susper

Games (Lapuschkin'18)



VQA (Arras'18)



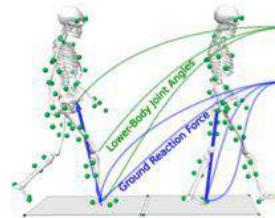
Video (Anders'18)



Morphing (Seibold'18)



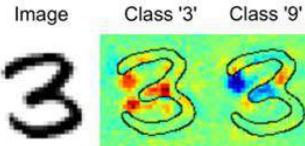
Gait Patterns (Horst'18)



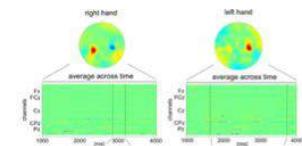
Faces (Lapuschkin'17)



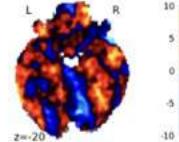
Digits (Bach' 15)



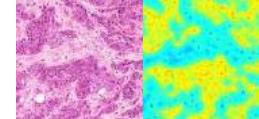
EEG (Sturm'16)



fMRI (Thomas'18)

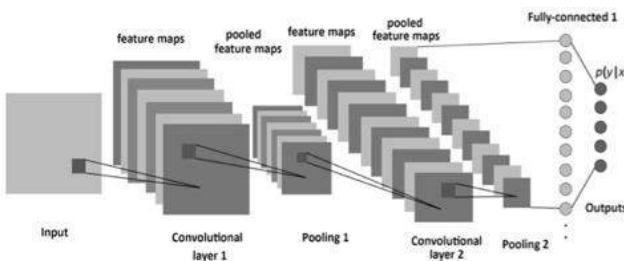


Histopathology (Binder'18)

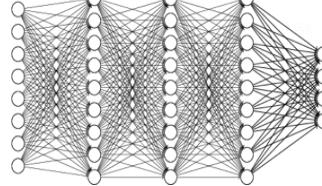


Summary LRP

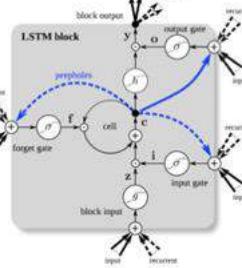
Convolutional NNs (Bach'15, Arras'17 ...)



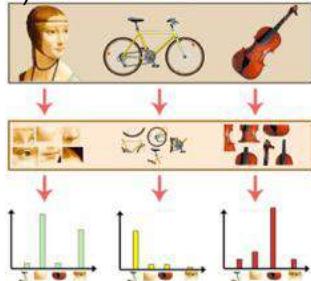
Local Renormalization Layers (Binder'16)



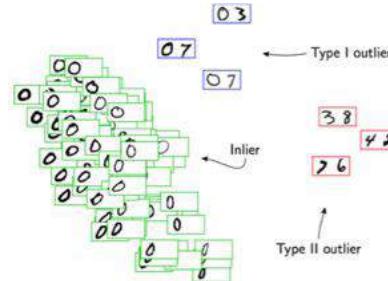
LSTM (Arras'17, Thomas'18)



Bag-of-words / Fisher Vector models
(Bach'15, Arras'16, Lapuschkin'17,
Binder'18)

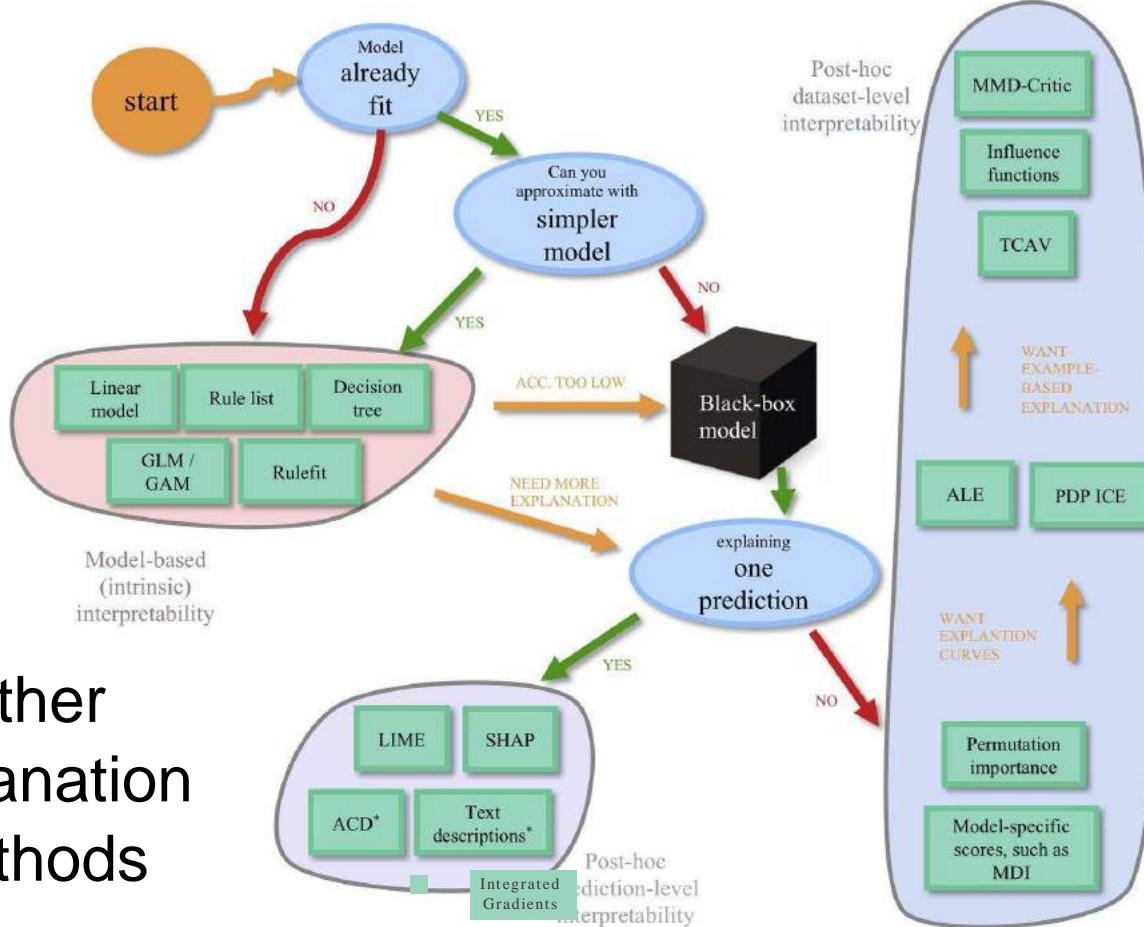


One-class SVM (Kauffmann'18)



Other Explanation Methods

* Denotes that a method only works on certain models (e.g. only neural networks)



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

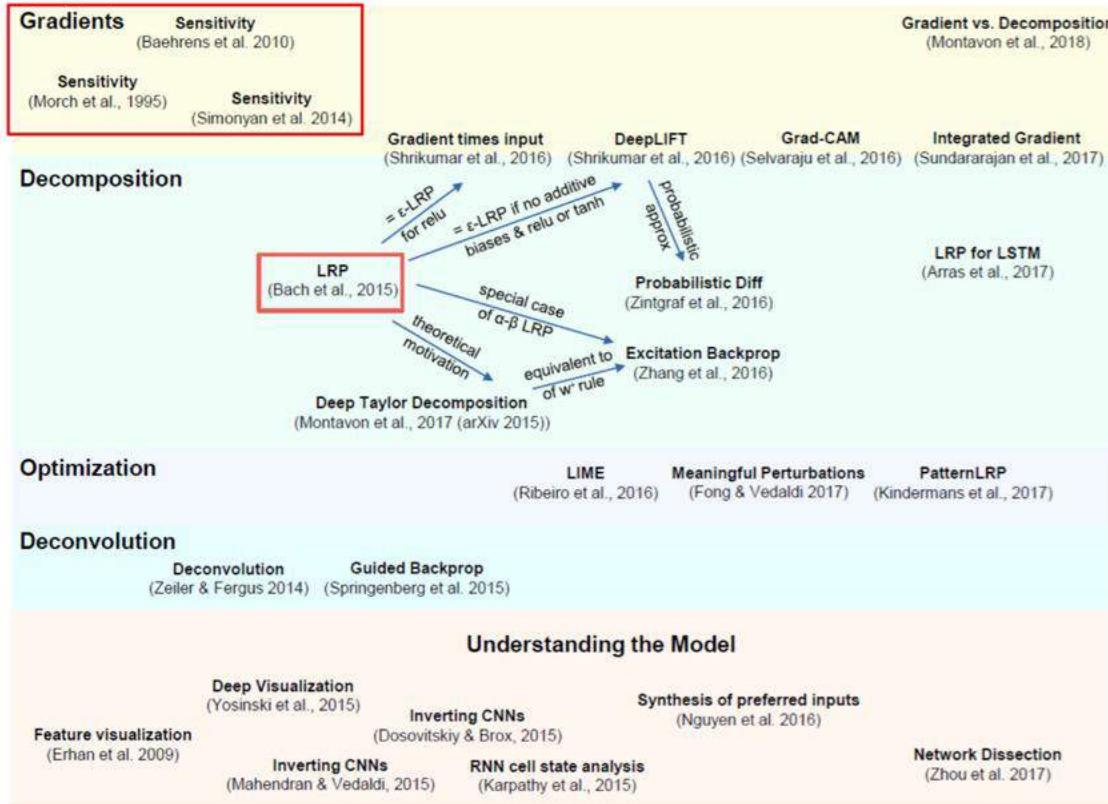
[PDP ICE](#) – vary feature value of all points and see how outputs change

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

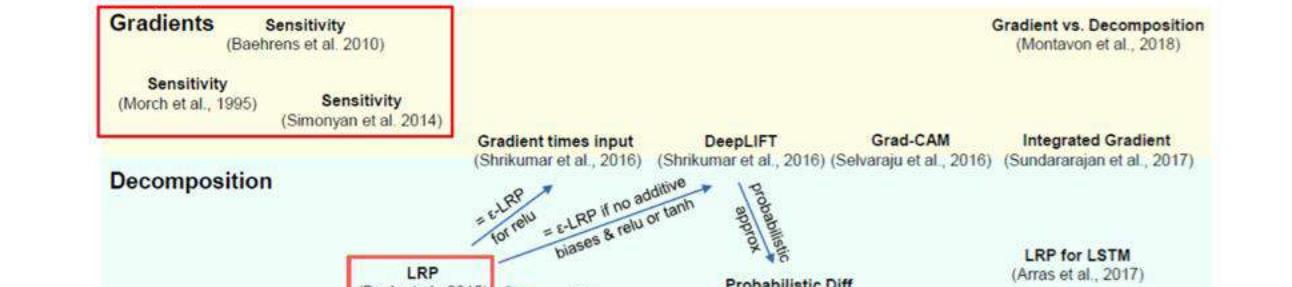
[Influence functions](#) – find points which highly influence a learned model

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

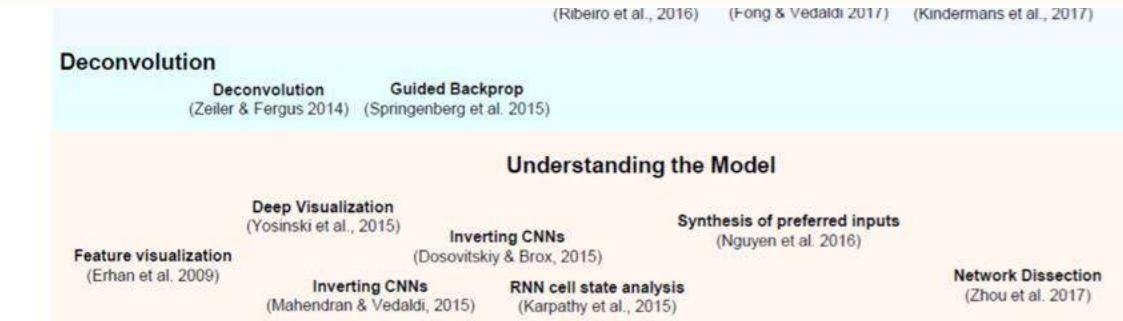
Other Explanation Methods



Other Explanation Methods

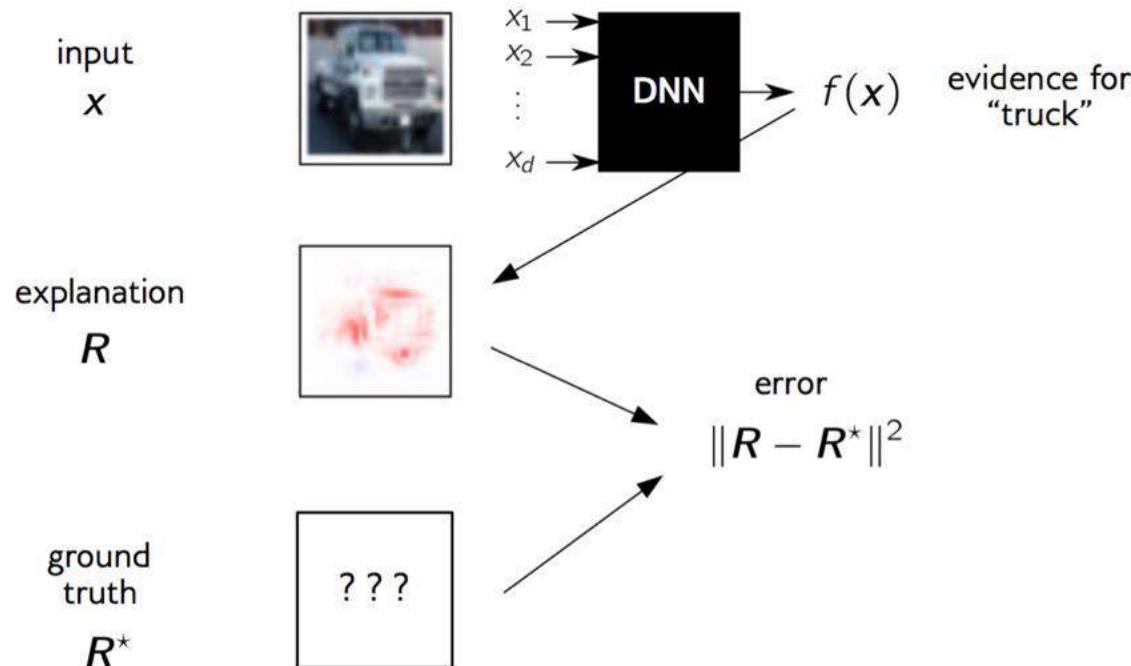


Question: Which one to choose ?



Axiomatic Approach to Interpretability

First Attempt: Distance to Ground Truth

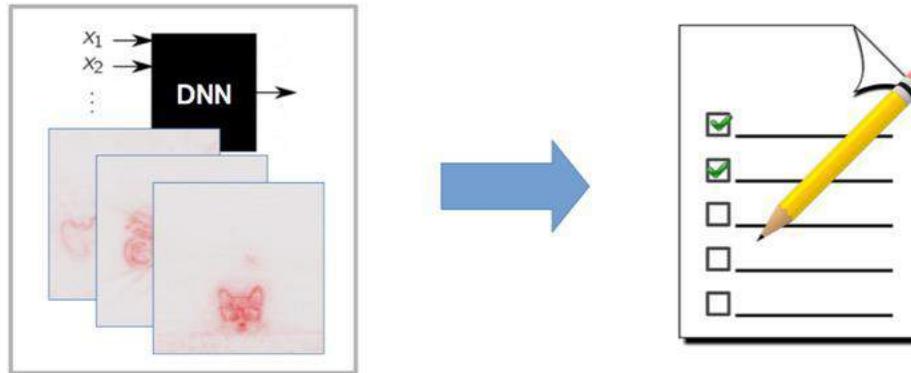


Axiomatic Approach to Interpretability

Idea: Evaluate the explanation technique axiomatically, i.e. it must pass a number of predefined “unit tests”.

[Sun’11, Bach’15, Montavon’17, Samek’17,
Sundarajan’17, Kindermans’17, Montavon’18].

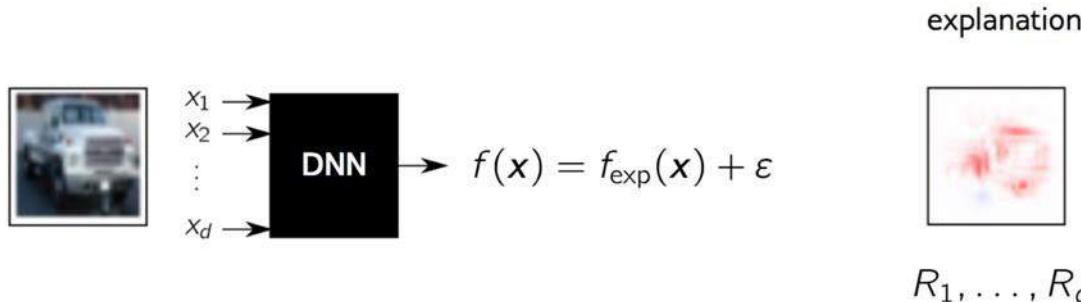
explanation technique



Axiomatic Approach to Interpretability

Properties 1-2: Conservation and Positivity

[Montavon'17, see also Sun'11, Landecker'13, Bach'15]



Conservation: Total attribution on the input features should be proportional to the amount of (explainable) evidence at the output.

$$\sum_{p=1}^d R_p = f_{\text{exp}}(x)$$

Positivity: If the neural network is certain about its prediction, input features are either relevant (positive) or irrelevant (zero).

$$\forall_{p=1}^d : R_p \geq 0$$

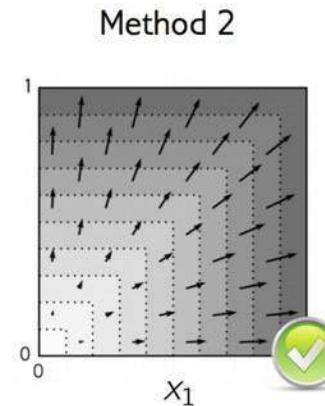
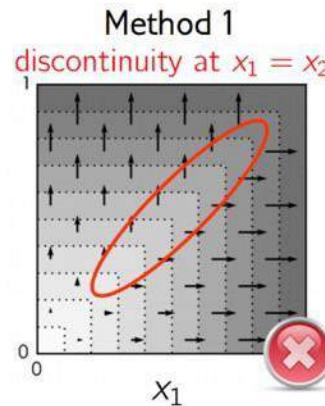
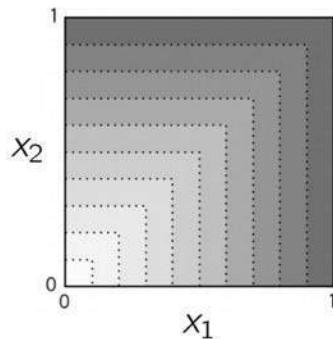
Axiomatic Approach to Interpretability

Property 3: Continuity [Montavon'18]

If two inputs are the almost the same, and the prediction is also almost the same, then the explanation should also be almost the same.

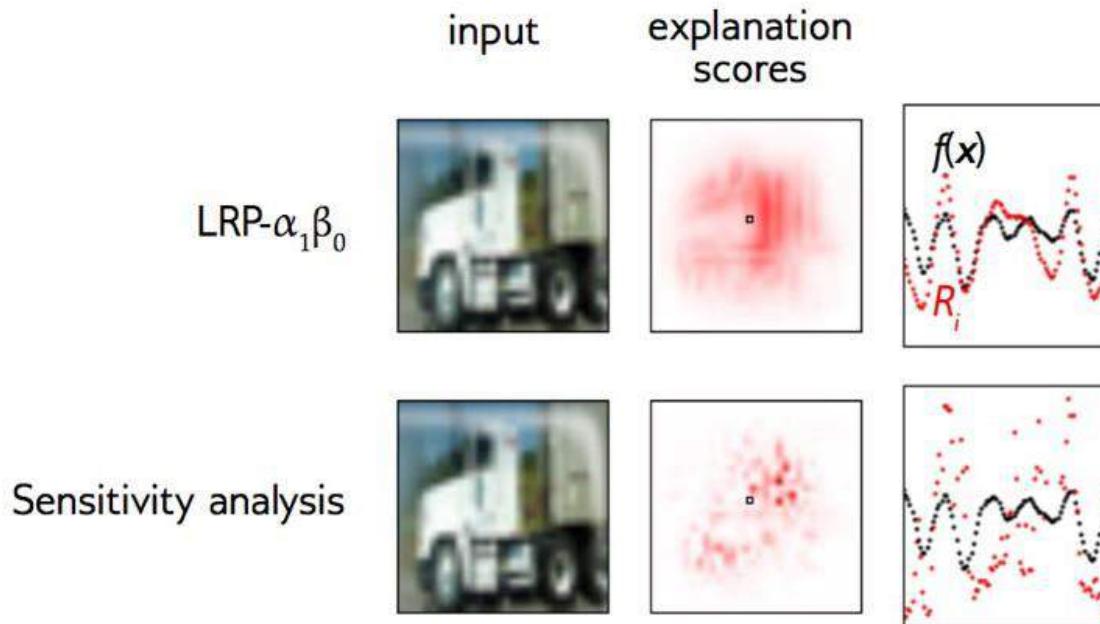
Example:

$$f(x) = \max(x_1, x_2)$$



Axiomatic Approach to Interpretability

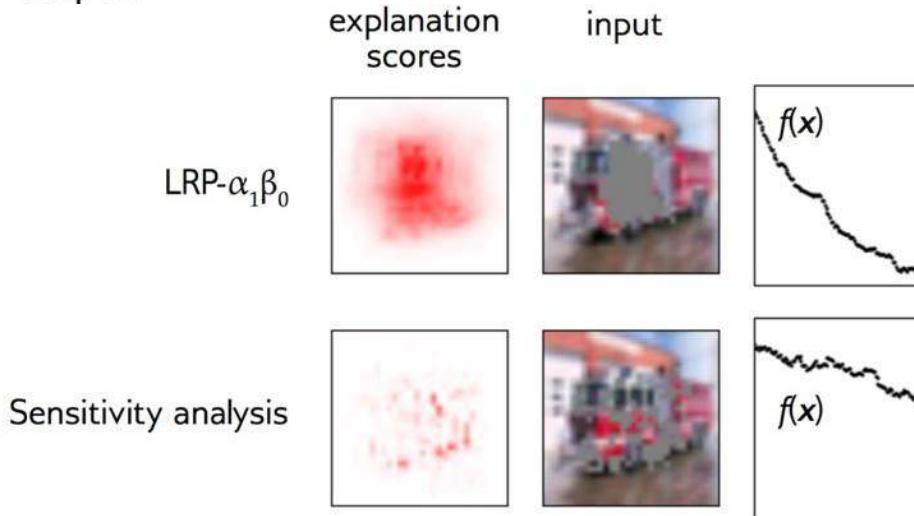
Testing Continuity



Axiomatic Approach to Interpretability

Property 4: Selectivity [Bach'15, Samek'17]

Model must agree with the explanation: If input features are attributed relevance, removing them should reduce evidence at the output.



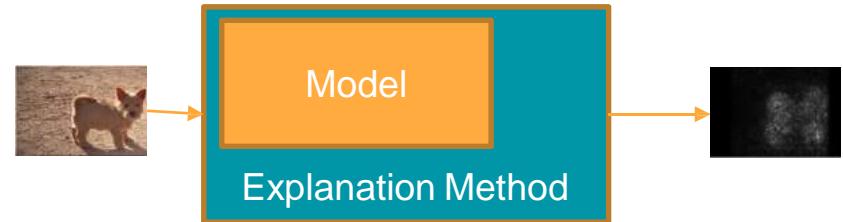
Axiomatic Approach to Interpretability

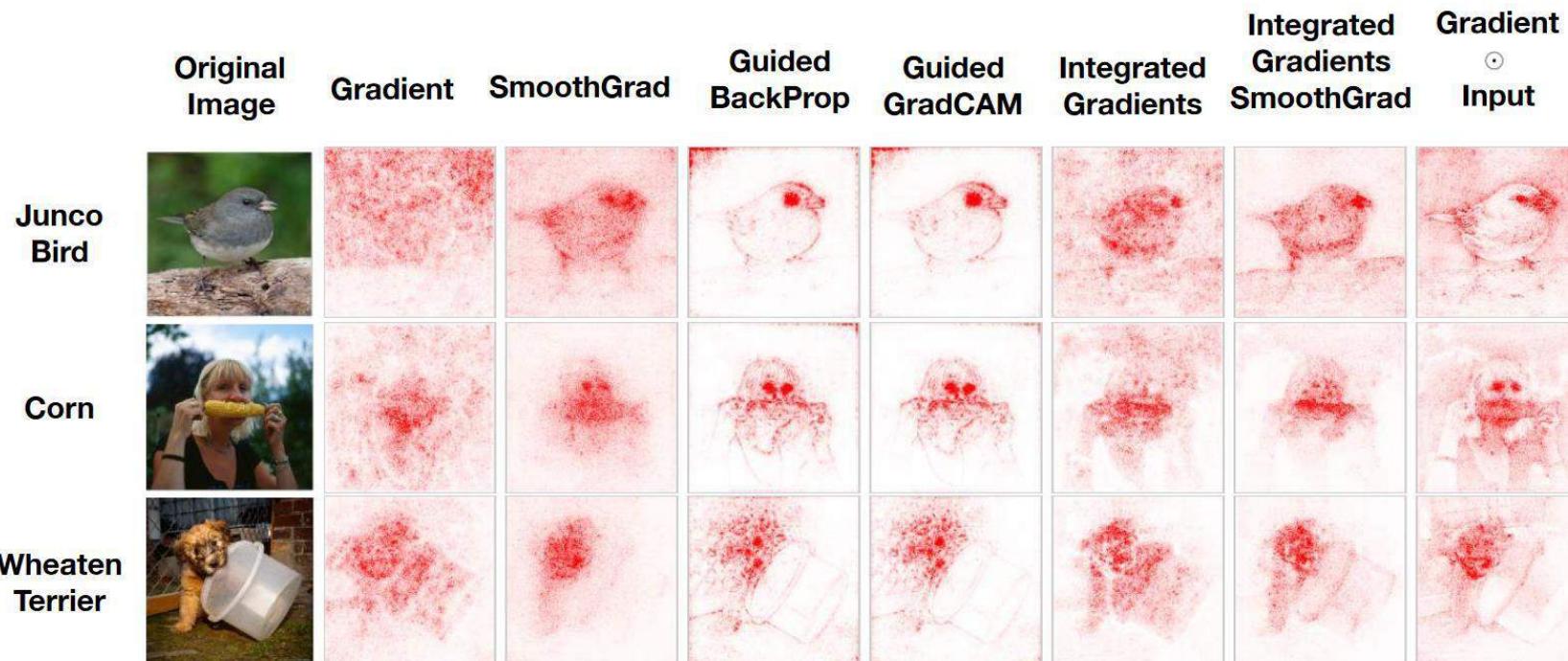
Explanation techniques	Uniform	$(\text{Gradient})^2$	$(\text{Guided BP})^2$	$\text{Gradient} \times \text{Input}$	$\text{Guided BP} \times \text{Input}$	$\text{LRP-}\alpha_i\beta_0$...
Properties							
1. Conservation	✓			✓	✓	✓	
2. Positivity	✓	✓	✓		✓	✓	
3. Continuity	✓		✓		✓	✓	
4. Selectivity		✓	✓	✓	✓	✓	
...							

Some Locally Interpretable, Post-hoc methods

Saliency Based Methods

- Heatmap based visualization
- Need differentiable model in most cases
- Normally involve gradient





[Adebayo et al 2018]

Saliency Example - Gradients

$$f(x): R^d \rightarrow R$$

$$E(f)(x) = \frac{df(x)}{dx}$$

How do we take gradient with respect to words?

Take gradient with respect to embedding of the word .

Saliency Example – Leave-one-out

$$f(x): R^d \rightarrow R$$

$$E(f)(x)_i = f(x) - f(x \setminus i)$$

How to remove ?

1. Zero out pixels in image
2. Remove word from the text
3. Replace the value with population mean in tabular data

Problems with Saliency Maps

- Only capture first order information
- Strange things can happen to heatmaps in second order.

[Feng et al 2018]

SQuAD

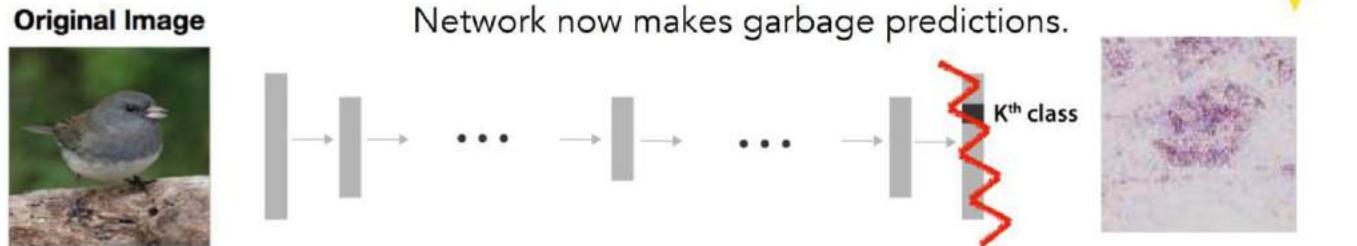
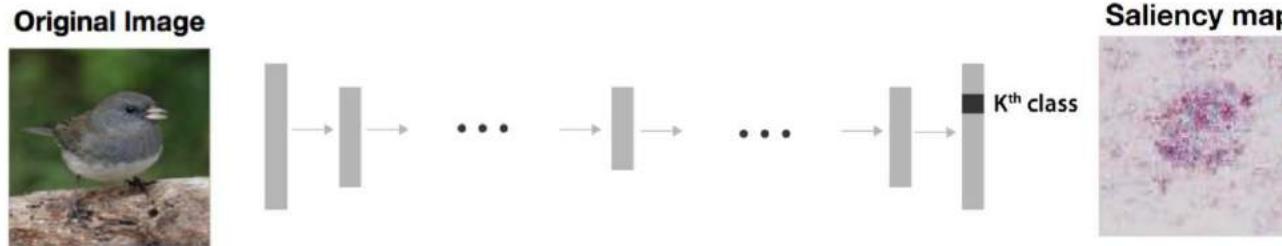
Context: QuickBooks sponsored a “Small Business Big Game” contest, in which Death Wish Coffee had a 30-second commercial aired free of charge courtesy of QuickBooks. [Death Wish Coffee](#) beat out nine other contenders from across the United States for the free advertisement.

Question:

What company won free advertisement due to QuickBooks contest ?
What company won free advertisement due to QuickBooks ?
What company won free advertisement due to ?
What company won free due to ?
What won free due to ?
What won due to ?
What won due to
What won due
What won
What

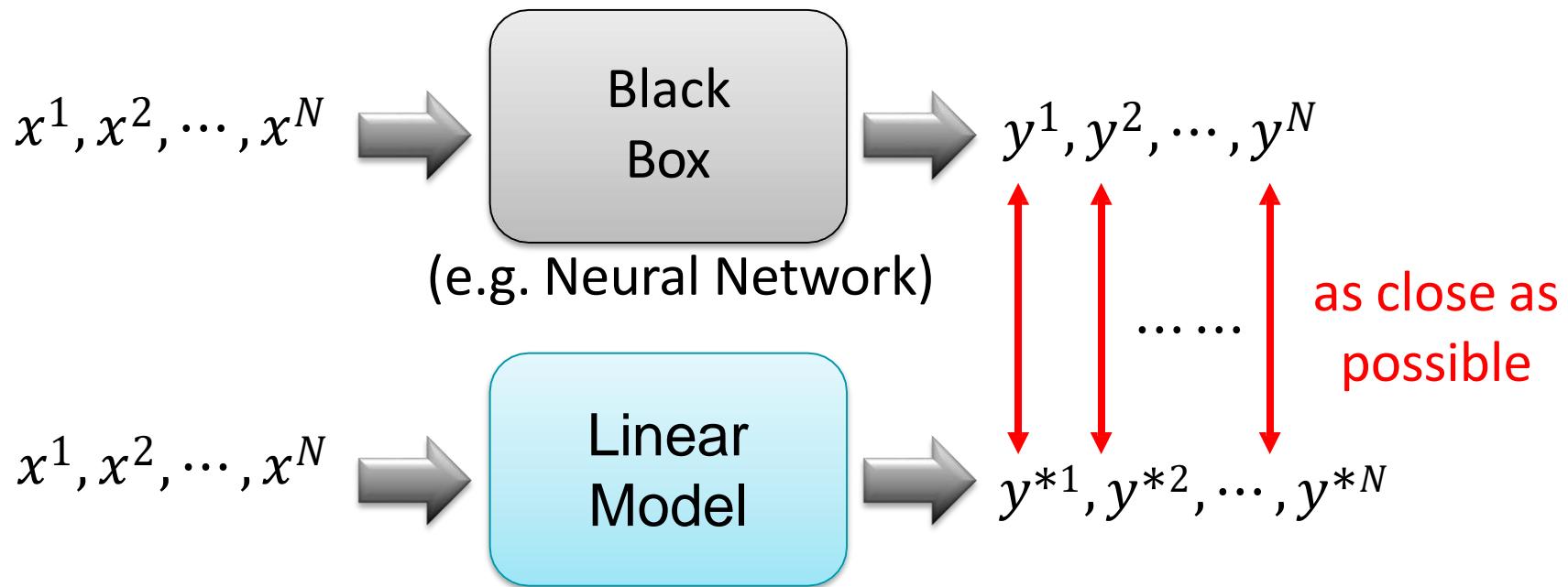
Figure 6: Heatmap generated with leave-one-out shifts drastically despite only removing the least important word (underlined) at each step. For instance, “advertisement”, is the most important word in step two but becomes the least important in step three.

Sanity check: When prediction changes, do explanations change?



(Slide Credit – Julius Adebayo)

LIME – locally interpretable model agnostic



Can't do it globally of course, but locally? Main Idea behind LIME

Intuition behind LIME

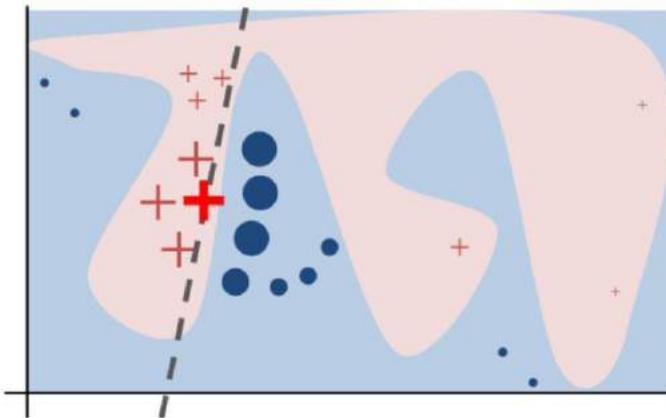
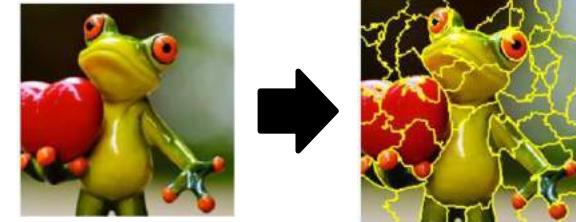


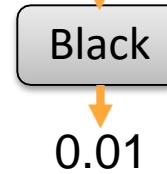
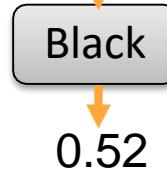
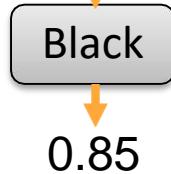
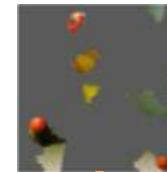
Figure 3: Toy example to present intuition for LIME. The black-box model's complex decision function f (unknown to LIME) is represented by the blue/pink background, which cannot be approximated well by a linear model. The bold red cross is the instance being explained. LIME samples instances, gets predictions using f , and weighs them by the proximity to the instance being explained (represented here by size). The dashed line is the learned explanation that is locally (but not globally) faithful.

[Ribeiro et al 2016]

LIME – Image



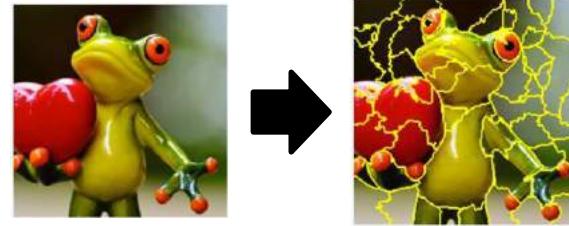
- Given a data point you want to explain
- Sample at the nearby - Each image is represented as a set of superpixels (segments).



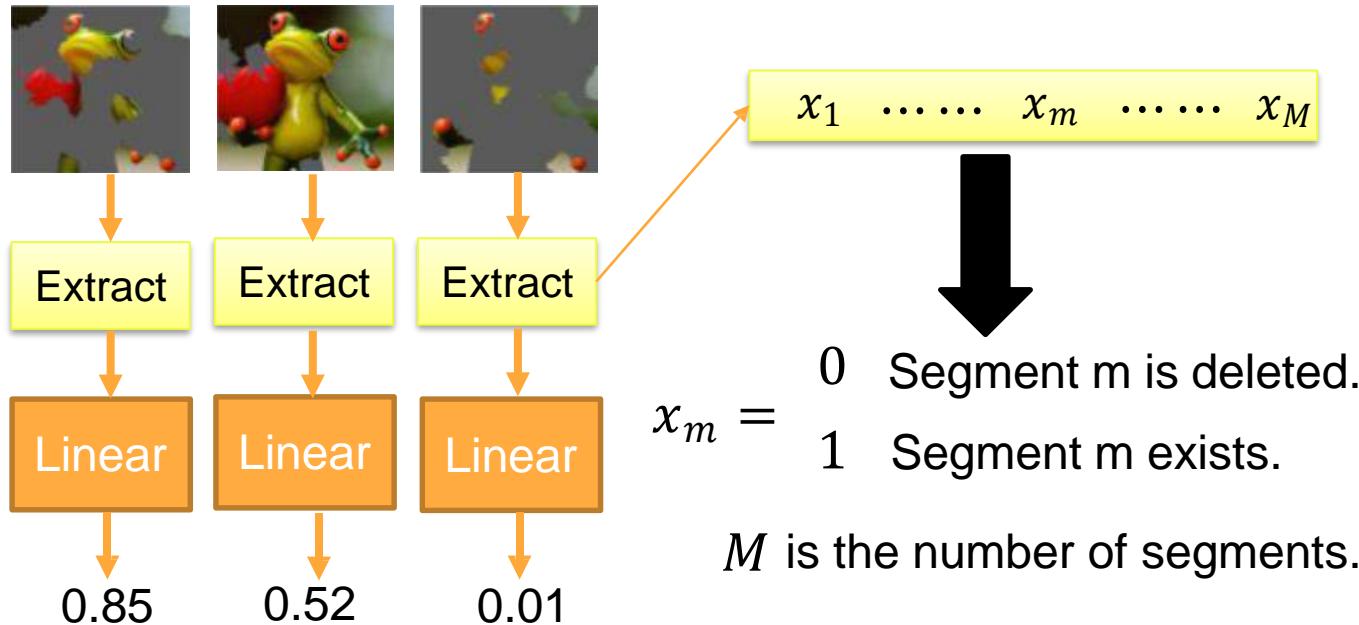
Randomly delete some segments.

Compute the probability of “frog” by black box

LIME – Image

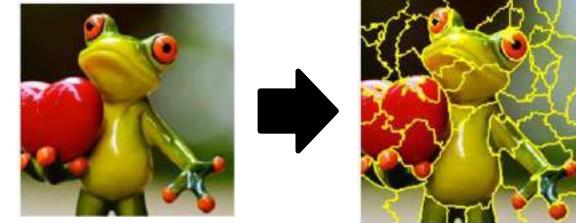


- 3. Fit with linear (or interpretable) model

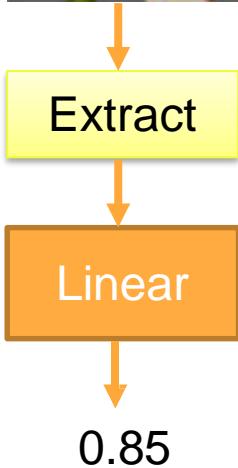
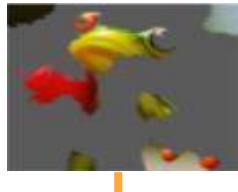


(Slide Credit – Hung-yi Lee)

LIME – Image



- 4. Interpret the model you learned



$$y = w_1 x_1 + \dots + w_m x_m + \dots + w_M x_M$$

$w_m = 0$ Segment m is deleted.

$x_m = 1$ Segment m exists.

M is the number of segments.

If $w_m \approx 0$ → segment m is not related to “frog”

If w_m is positive → segment m indicates the image is “frog”

If w_m is negative → segment m indicates the image is not “frog”

The Math behind LIME

Algorithm 1 Sparse Linear Explanations using LIME

Require: Classifier f , Number of samples N

Require: Instance x , and its interpretable version x'

Require: Similarity kernel π_x , Length of explanation K

```
 $\mathcal{Z} \leftarrow \{\}$ 
for  $i \in \{1, 2, 3, \dots, N\}$  do
     $z'_i \leftarrow \text{sample\_around}(x')$ 
     $\mathcal{Z} \leftarrow \mathcal{Z} \cup \langle z'_i, f(z_i), \pi_x(z_i) \rangle$ 
end for
```

Match interpretable
model to black box

Control
complexity of the
model

$$\xi(x) = \operatorname{argmin}_{g \in G} \mathcal{L}(f, g, \pi_x) + \Omega(g)$$

$$\mathcal{L}(f, g, \pi_x) = \sum_{z, z' \in \mathcal{Z}} \pi_x(z) (f(z) - g(z'))^2$$

Rationalization Models

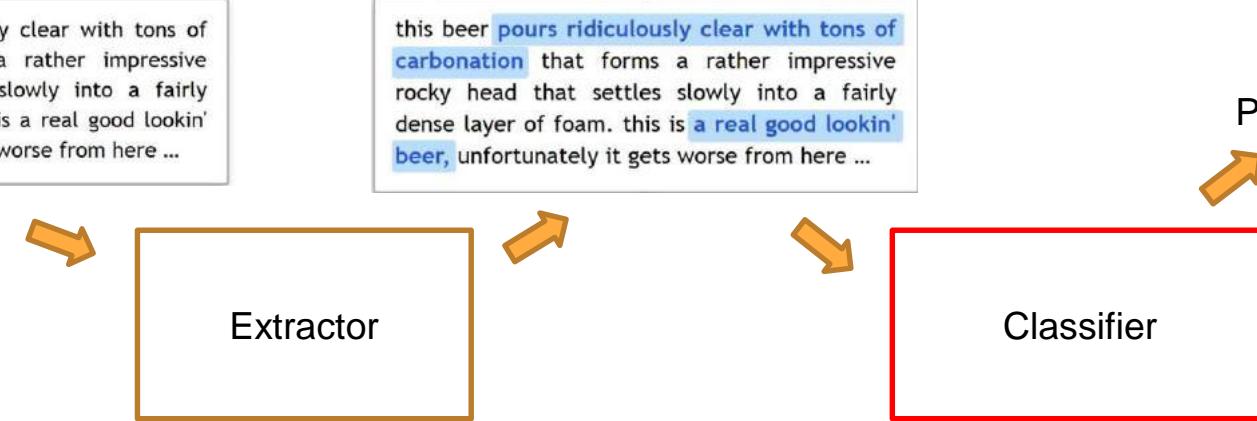
General Idea



this beer pours ridiculously clear with tons of carbonation that forms a rather impressive rocky head that settles slowly into a fairly dense layer of foam. this is a real good lookin' beer, unfortunately it gets worse from here ...

this beer **pours ridiculously clear with tons of carbonation** that forms a rather impressive rocky head that settles slowly into a fairly dense layer of foam. this is **a real good lookin' beer**, unfortunately it gets worse from here ...

Positive (98%)





Massachusetts
Institute of
Technology



Rationalizing Neural Predictions

Tao Lei

Regina Barzilay Tommi Jaakkola

EMNLP 2016

(Slides Credit – Tao Lei)

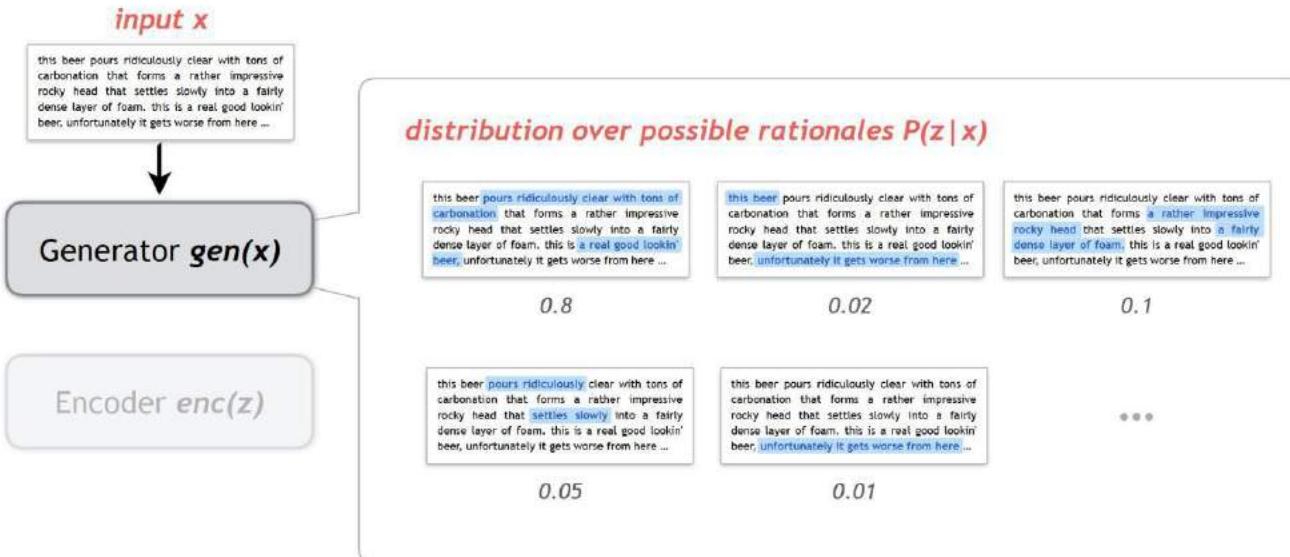
Model Architecture

Generator $\mathbf{gen}(\mathbf{x})$

Encoder $\mathbf{enc}(\mathbf{z})$

two modular components $\mathbf{gen}()$ and $\mathbf{enc}()$

Model Architecture

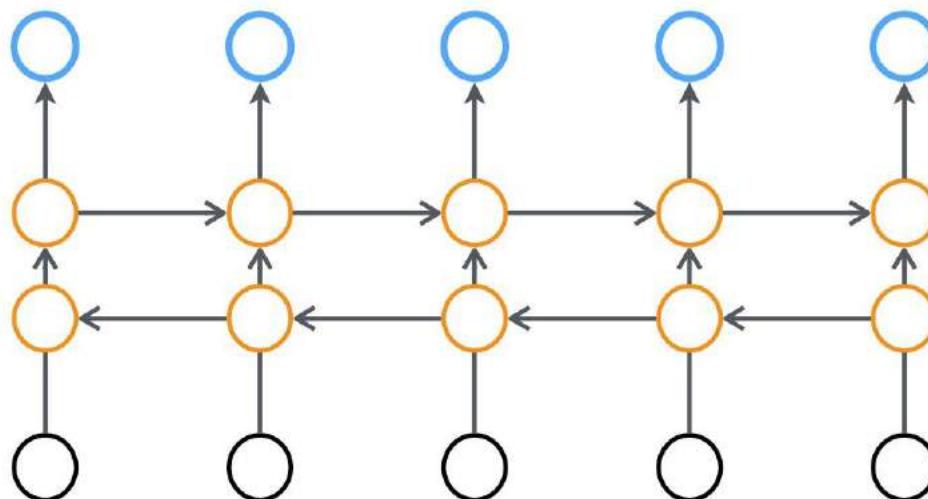


generator specifies the distribution of rationales

binary selection z :

0 1 0 1 1

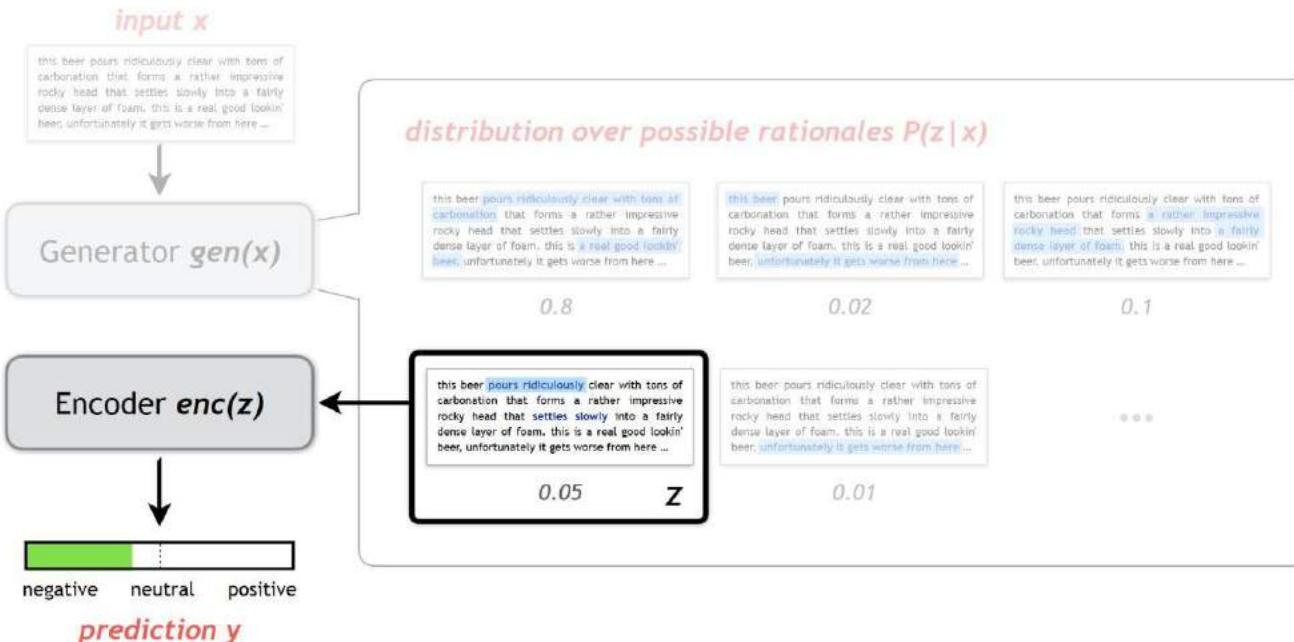
$P(z)$:



hidden states:

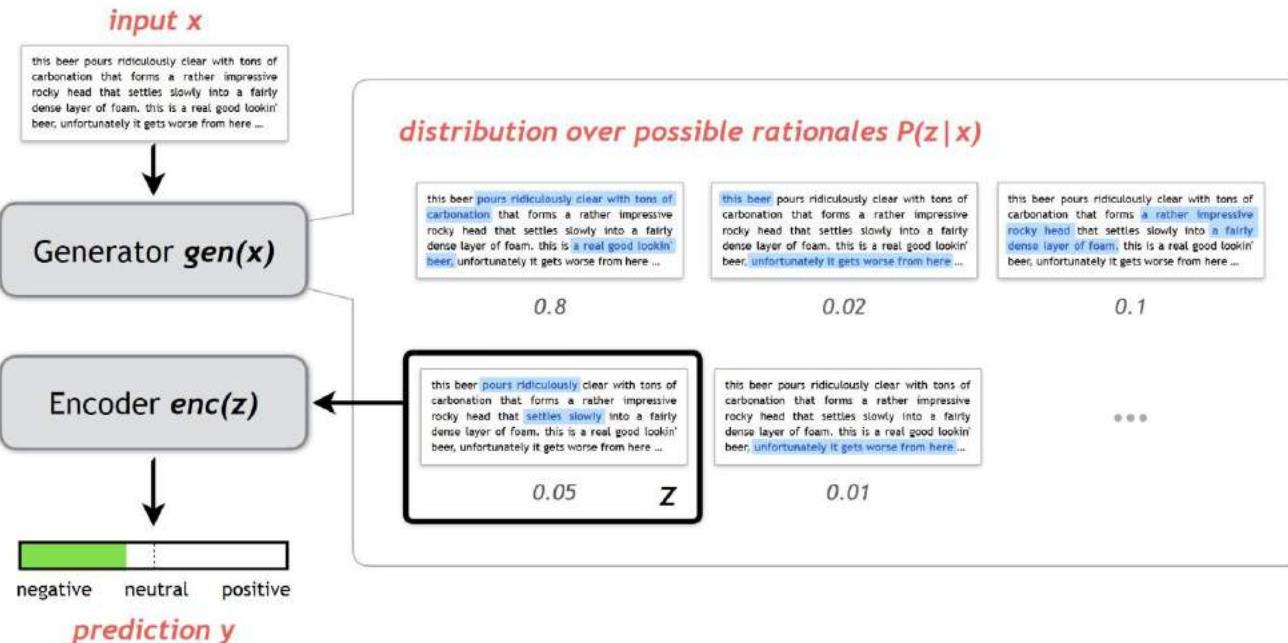
input words x :

Model Architecture



encoder makes prediction given rationale

Model Architecture



Training Objective

$$\text{cost}(\mathbf{z}, \mathbf{y}) = \text{loss}(\mathbf{z}, \mathbf{y}) + \lambda_1 |\mathbf{z}|_1 + \lambda_2 \sum_i |\mathbf{z}_i - \mathbf{z}_{i-1}|$$

sufficiency
correct prediction

sparsity
rationale is short

coherency
continuous selection

- receive this training signal after \mathbf{z} is produced

Minimizing expected cost:

$$\min_{\theta} \sum_{(\mathbf{x}, \mathbf{y}) \in D} \mathbb{E}_{\mathbf{z} \sim \text{gen}(\mathbf{x})} [\text{cost}(\mathbf{z}, \mathbf{y})]$$

- intractable because summation over \mathbf{z} is exponential

Learning Method

- Possible to sample the gradient, e.g.:

$$\mathbb{E}_{\mathbf{z} \sim \text{gen}(\mathbf{x})} \left[\text{cost}(\mathbf{z}, \mathbf{y}) \frac{\partial \log P(\mathbf{z}|\mathbf{x})}{\partial \theta_g} \right]$$

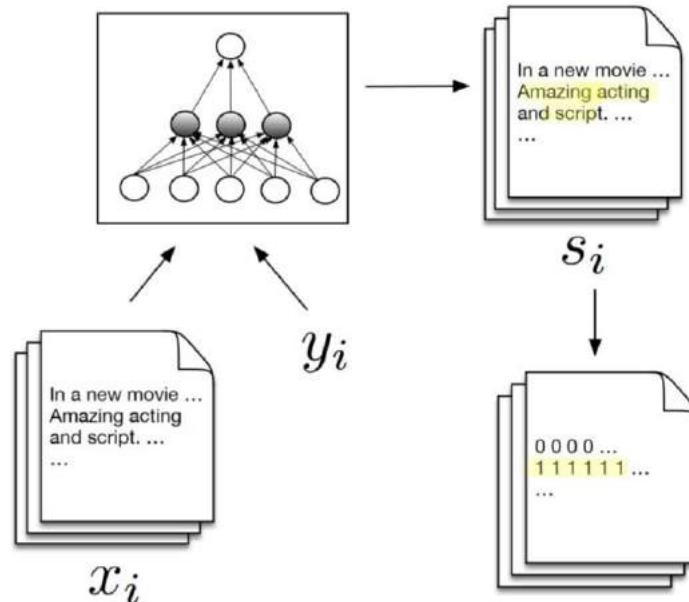
$$\approx \frac{1}{N} \sum_{i=1}^N \text{cost}(\mathbf{z}_i, \mathbf{y}_i) \frac{\partial \log P(\mathbf{z}_i|\mathbf{x}_i)}{\partial \theta_g}$$

where \mathbf{z}_i are sampled rationales

- Stochastic gradient decent on sampled gradients

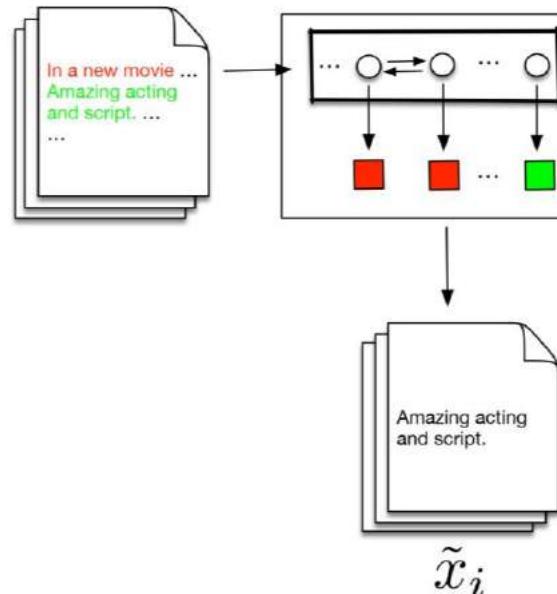
FRESH Model – Faithful Rationale Extraction using Saliency Thresholding

(1) Train supp to score features (e.g., gradients, attention, LIME); discretize these



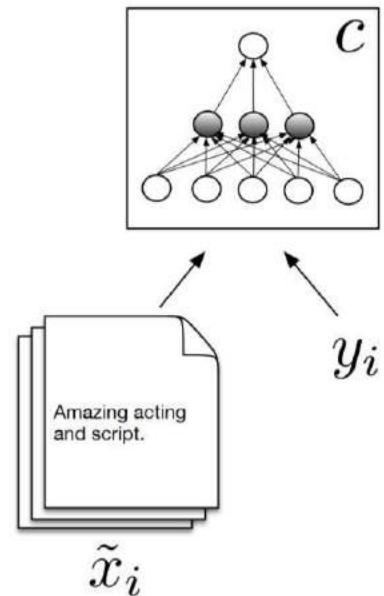
FRESH Model – Faithful Rationale Extraction using Saliency Thresholding

(2) *Train ext to extract snippets; use to create \tilde{x}_i*



FRESH Model – Faithful Rationale Extraction using Saliency Thresholding

(3) Train pred on (\tilde{x}_i, y_i)



From LRP to Deep Taylor Decomposition

Summary LRP

1. LRP solves the “correct” explanation problem
2. It has a theoretical interpretation (Deep Taylor Decomposition)
3. It can be applied to various data and models (not only deep nets)
4. It fulfills various criteria (axiomatic approach)
5. It is flexible (many explanation methods are special cases of LRP)
6. In general: $\text{LRP} \neq \text{Gradient} \times \text{Input}$

Tutorial Paper

Montavon et al., “Methods for interpreting and understanding deep neural networks”,
Digital Signal Processing, 73:1-5, 2018

56

Keras Explanation Toolbox

<https://github.com/albermax/investigate>

Decomposing the Correct Quantity

slope decomposition

$$\sum_i R_i = \|\nabla_x f\|^2$$

value decomposition

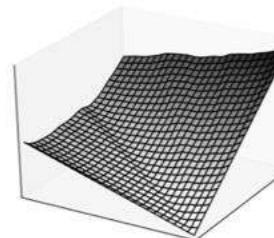
$$\sum_i R_i = f(\mathbf{x})$$

Candidate: Taylor decomposition

$$f(\mathbf{x}) = \underbrace{f(\tilde{\mathbf{x}})}_0 + \sum_{i=1}^d \underbrace{\frac{\partial f}{\partial x_i} \Big|_{\mathbf{x}=\tilde{\mathbf{x}}} (x_i - \tilde{x}_i)}_{R_i} + \underbrace{O(\mathbf{x}\mathbf{x}^\top)}_0$$

- ▶ Achievable for linear models and deep ReLU networks without biases, by choosing:

$$\tilde{\mathbf{x}} = \lim_{\varepsilon \rightarrow 0} \varepsilon \cdot \mathbf{x} \approx \mathbf{0}.$$



Why Simple Taylor doesn't work?

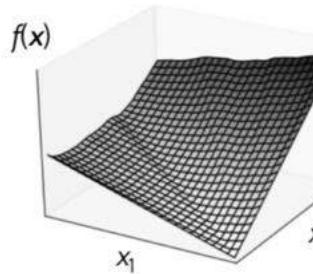
Two Reasons:

1

Root point is hard to find or too far → includes too much information (incl. negative evidence)

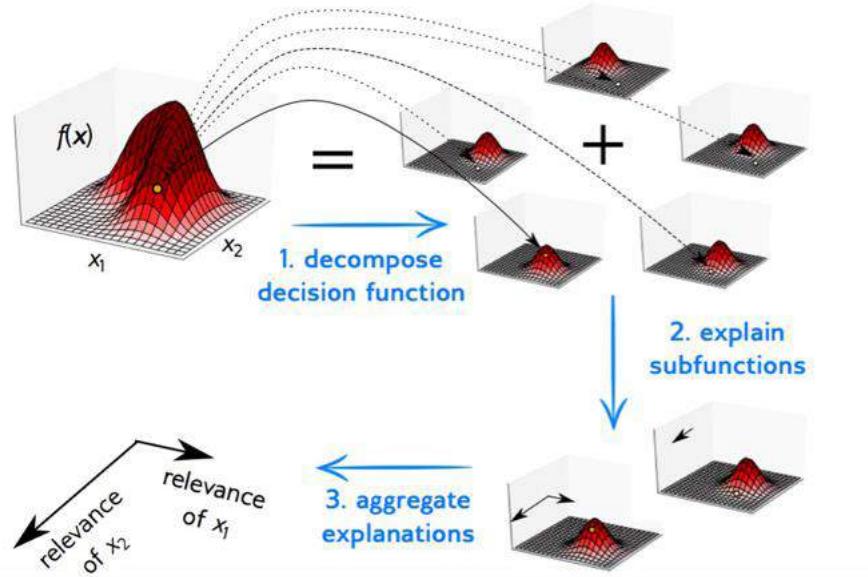
2

Gradient shattering problem → gradient of deep nets has low informative value



Deep Taylor Decomposition

Idea: Since neural network is composed of simple functions, we propose a *deep* Taylor decomposition.

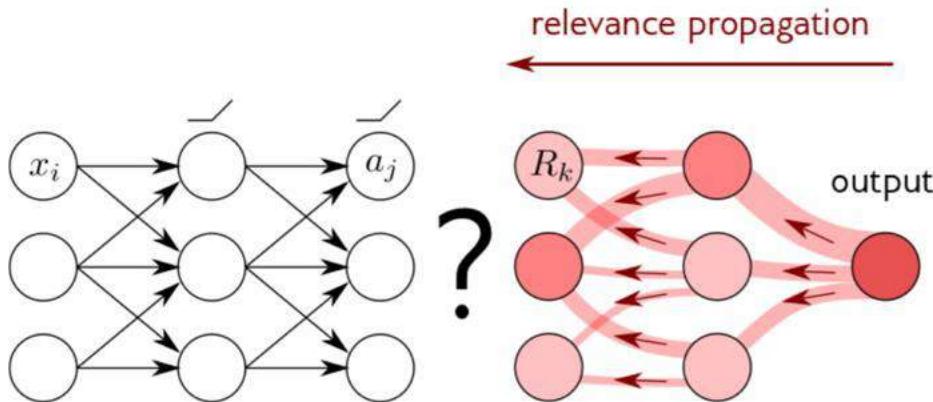


Each explanation step:

- easy to find good root point
- no gradient shattering

(Montavon et al., 2017
Montavon et al. 2018)

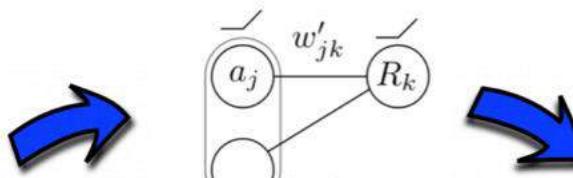
Deep Taylor Decomposition



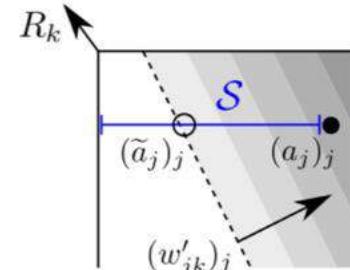
Can we express R_k as a simple function of $(a_j)_j$?

Can we do a Taylor decomposition of $R_k((a_j)_j)$?

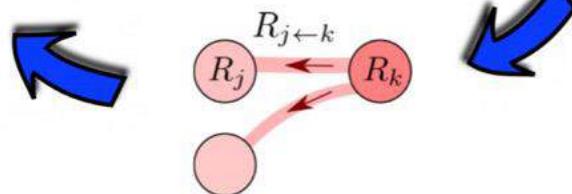
Deep Taylor Decomposition



Observe that $R_k \approx a_k \cdot \text{const.}$

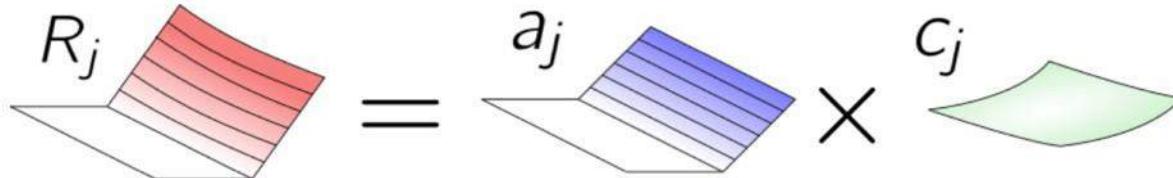


Move to the lower-layer



Deep Taylor Decomposition

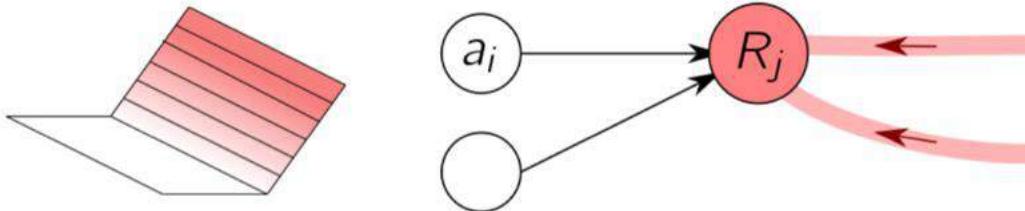
Proposition: Relevance at each layer is a product of the activation and an approximately constant term.

$$R_j = a_j \times C_j$$


Deep Taylor Decomposition

1

Build the Relevance Neuron



$$R_j = a_j c_j$$

$$= \max(0, \sum_i a_i w_{ij}) \cdot c_j$$

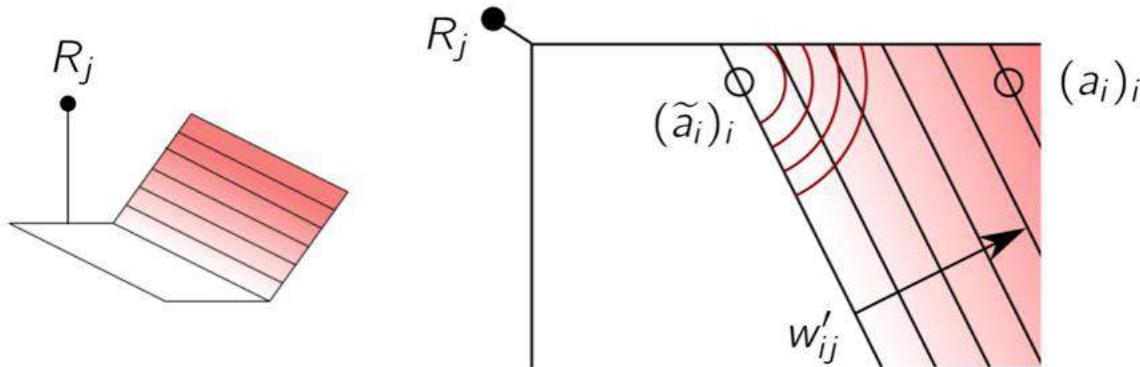
$$= \max(0, \sum_i a_i w'_{ij})$$

$$w'_{ij} = w_{ij} c_j$$

Deep Taylor Decomposition

2 Expand the Relevance Neuron

$$R_j((a_i)_i) = R_j((\tilde{a}_i)_i) + \sum_i \underbrace{\frac{\partial R_j}{\partial a_i} \Big|_{(\tilde{a}_i)_i} \cdot (a_i - \tilde{a}_i)}_{R_{i \leftarrow j}} + \varepsilon$$



Deep Taylor Decomposition

3

Decompose Relevance

Taylor expansion at root point:

$$R_j(\mathbf{a}) = R_j(\tilde{\mathbf{a}}^{(j)}) + \sum_i \frac{\partial R_j}{\partial a_i} \Big|_{\tilde{\mathbf{a}}^{(j)}} \cdot (a_i - \tilde{a}_i^{(j)}) + \varepsilon$$

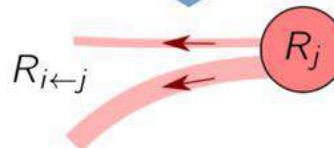
$$\downarrow \qquad \qquad \qquad \downarrow \qquad \qquad \qquad \downarrow$$

0

$$\frac{(a_i - \tilde{a}_i^{(j)}) w_{ij}}{\sum_i (a_i - \tilde{a}_i^{(j)}) w_{ij}} R_j$$

0

Relevance can now be
backward propagated



Deep Taylor Decomposition

$$R_{i \leftarrow j} = \frac{(a_i - \tilde{a}_i^{(j)})w_{ij}}{\sum_i (a_i - \tilde{a}_i^{(j)})w_{ij}} R_j \quad (\text{Deep Taylor generic})$$



Choice of root point

	$\tilde{a}^{(j)} \in \mathcal{D}$	$\ a - \tilde{a}^{(j)}\ $
1. nearest root	$\tilde{a}^{(j)} = a - t \cdot w_j$	✓
2. rescaled activation	$\tilde{a}^{(j)} = a - t \cdot a$	✓
3. rescaled excitations	$\tilde{a}^{(j)} = a - t \cdot a \odot \mathbf{1}_{w_j > 0}$	✓ ✓



$$R_{i \leftarrow j} = \frac{a_i w_{ij}^+}{\sum_i a_i w_{ij}^+} R_j \quad (\text{LRP-}\alpha_1\beta_0)$$

Deep Taylor Decomposition

Input domain	Rule
ReLU activations $(a_j \geq 0)$	$R_j = \sum_k \frac{a_j w_{jk}^+}{\sum_j a_j w_{jk}^+} R_k$
Pixel intensities $(x_i \in [l_i, h_i], l_i \leq 0 \leq h_i)$	$R_i = \sum_j \frac{x_i w_{ij} - l_i w_{ij}^+ - h_i w_{ij}^-}{\sum_i x_i w_{ij} - l_i w_{ij}^+ - h_i w_{ij}^-} R_j$
Real values $(x_i \in \mathbb{R})$	$R_i = \sum_j \frac{w_{ij}^2}{\sum_i w_{ij}^2} R_j$

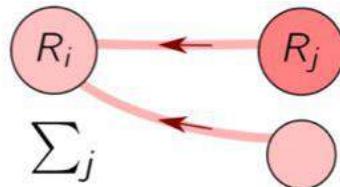
Deep Taylor LRP rules [Montavon'17]

More refined rules can also be constructed to match the input data distribution [Kindermans'17]

Deep Taylor Decomposition

4

Pooling relevance
over all outgoing
neurons



Deep Taylor Decomposition

The LRP- $\alpha_1\beta_0$ rule

$$R_i = \sum_j \frac{a_i w_{ij}^+}{\sum_i a_i w_{ij}^+} R_j$$

can be seen as

a deep Taylor
decomposition (DTD)

[Montavon'17]

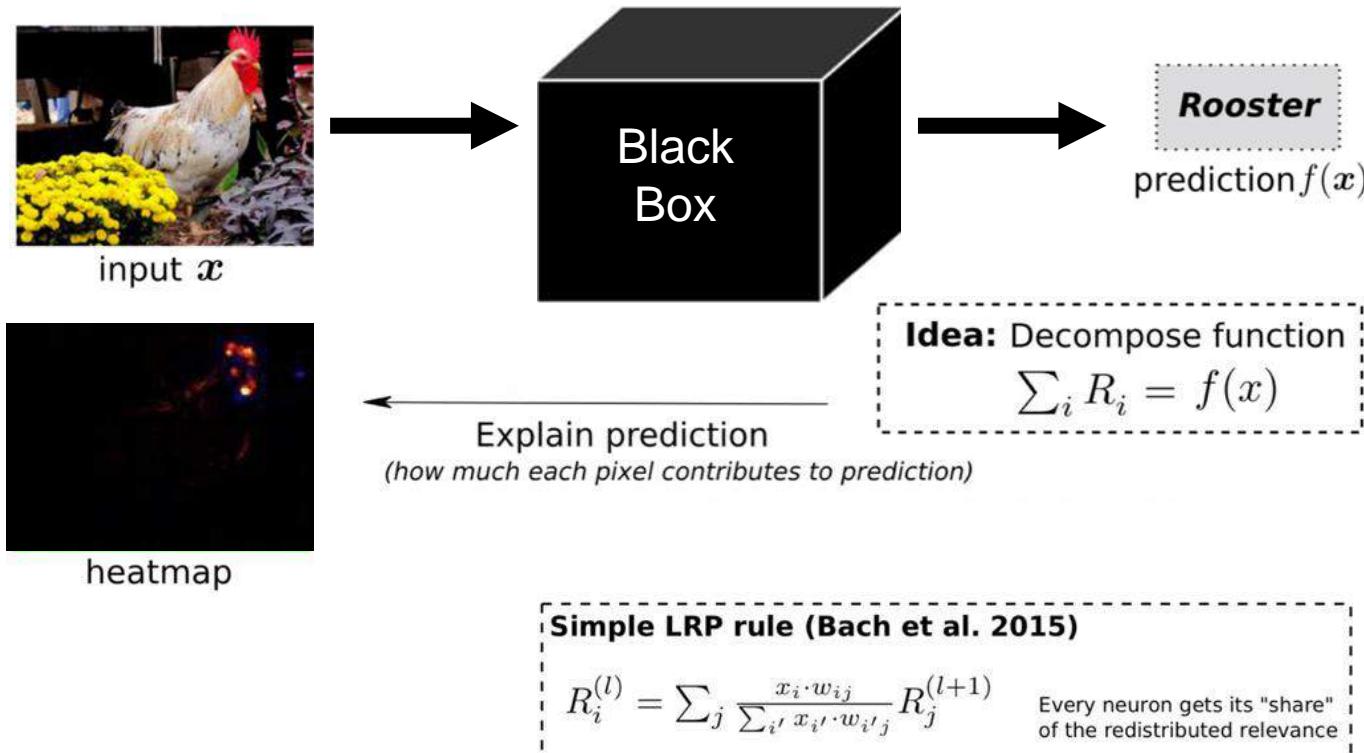
which then yields

domain- and layer-
specific rules

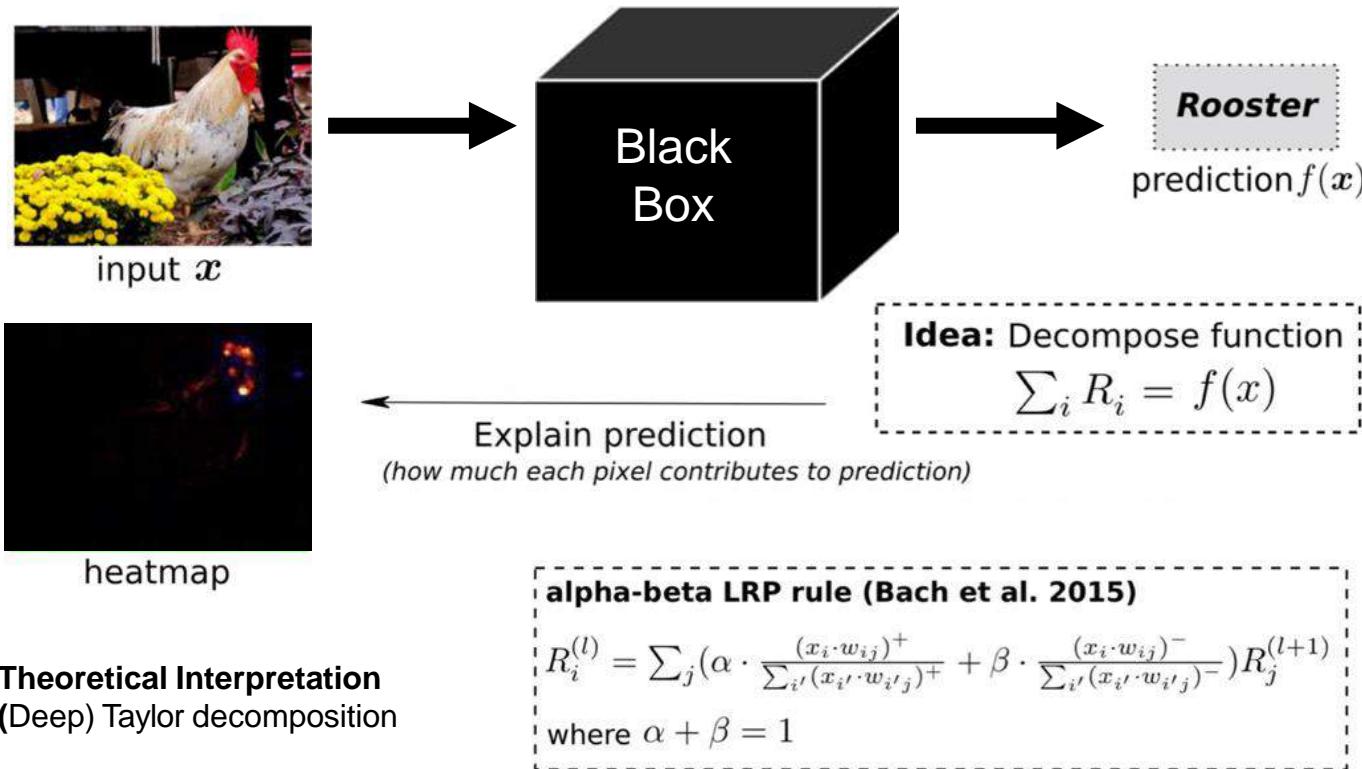
Tutorial on Interpretable Machine Learning

Part 3: Applications of Interpretability

LRP revisited



LRP revisited

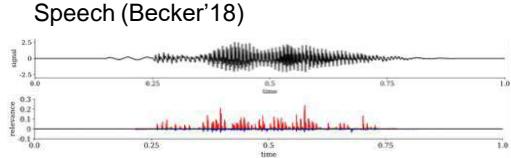


LRP revisited

General Images (Bach' 15, Lapuschkin'16)



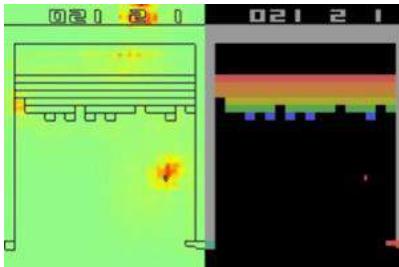
Speech (Becker'18)



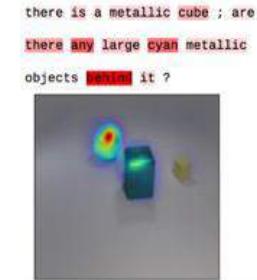
Text Analysis (Arras'16 &17)

do n't waste your money
neither funny nor susper

Games (Lapuschkin'18)



VQA (Arras'18)



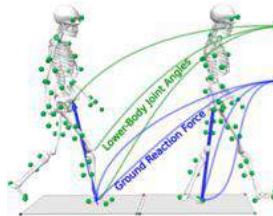
Video (Anders'18)



Morphing (Seibold'18)



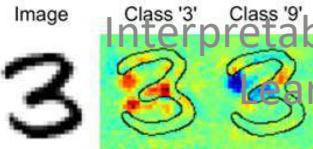
Gait Patterns (Horst'18)



Faces (Lapuschkin'17)



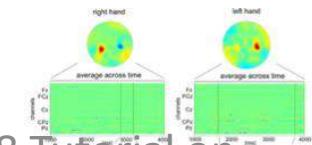
Digits (Bach' 15)



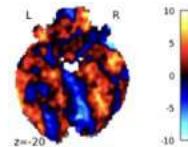
MICCAI'18 Tutorial on
Interpretable Machine
Learning



EEG (Sturm'16)

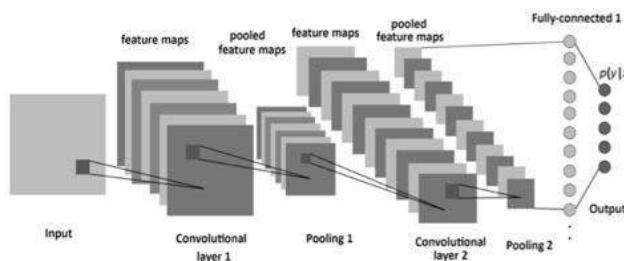


fMRI (Thomas'18) 74

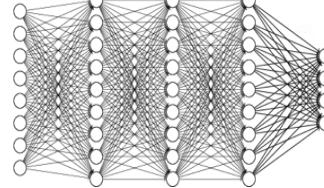


LRP revisited

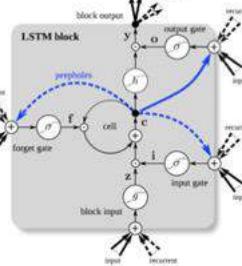
Convolutional NNs (Bach'15, Arras'17 ...)



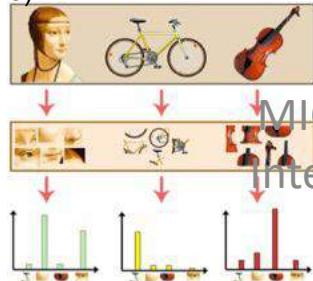
Local Renormalization Layers (Binder'16)



LSTM (Arras'17, Thomas'18)

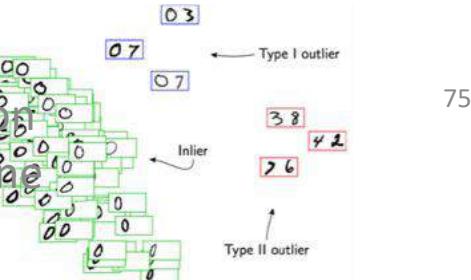


Bag-of-words / Fisher Vector models
(Bach'15, Arras'16, Lapuschkin'17,
Binder'18)



MICCAI'18 Tutorial
Interpretable Machine
Learning

One-class SVM (Kauffmann'18)



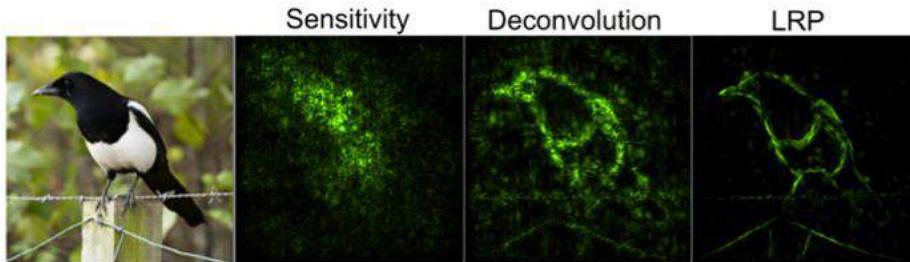
75

LRP & Others

Evaluating Heatmap Quality

MICCAI'18 Tutorial on
Interpretable Machine
Learning

Compare Explanation Methods



Can we objectively measure which heatmap is best ?

Idea: Compare selectivity (Bach'15, Samek'17):

"If input features are deemed relevant, removing them should reduce evidence at the output of the network."

Algorithm (“Pixel Flipping”)

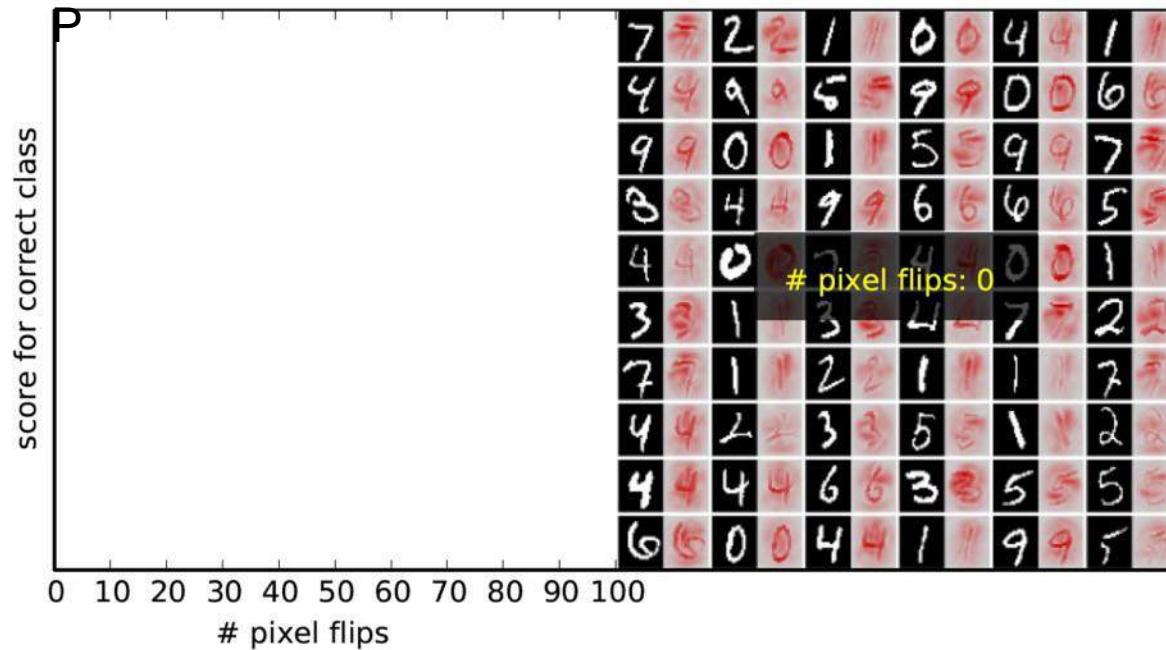
```
Sort pixels / patches by relevance  
Iterate  
    destroy pixel / patch  
    evaluate  $f(x)$   
Measure decrease of  $f(x)$ 
```

77

Important: Remove information in a non-specific manner (e.g. sample from uniform distribution)

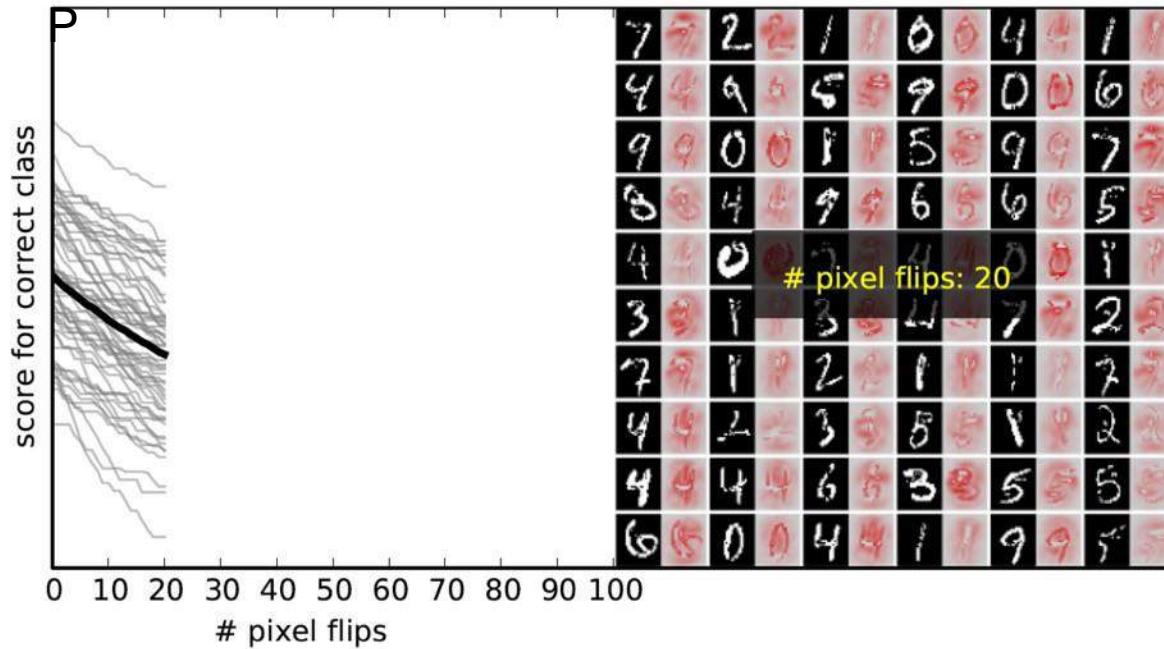
Compare Explanation Methods

LR

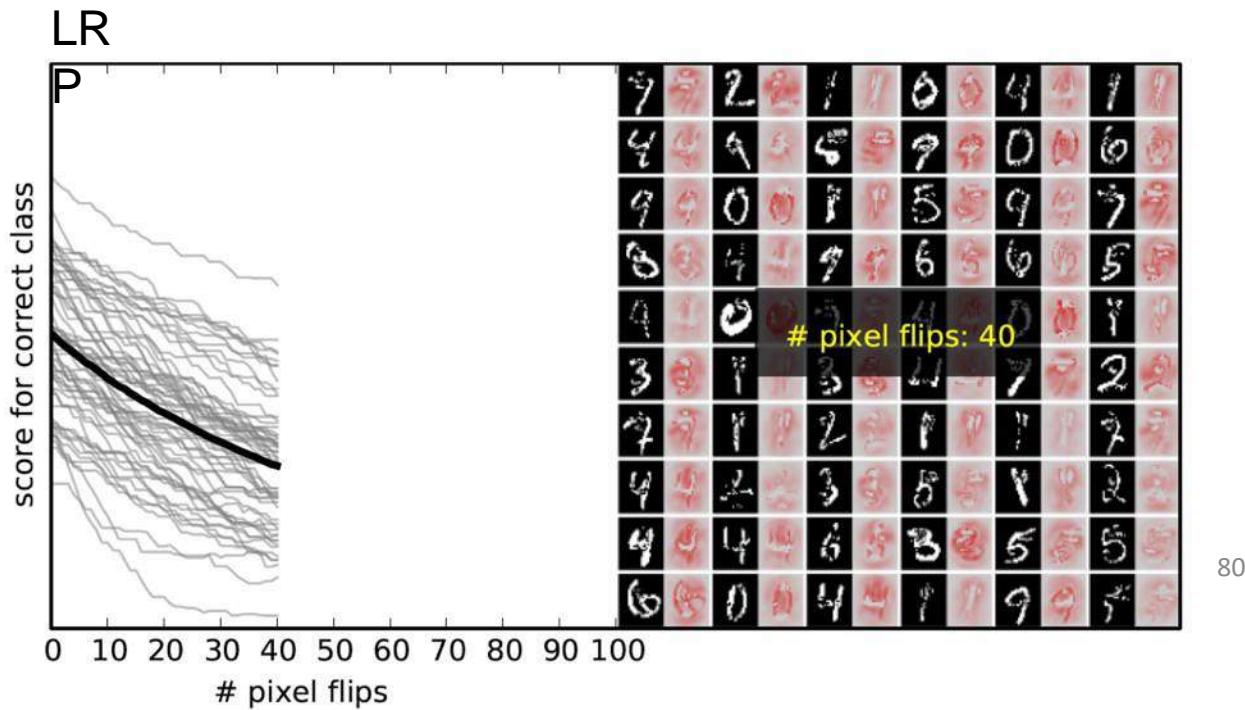


Compare Explanation Methods

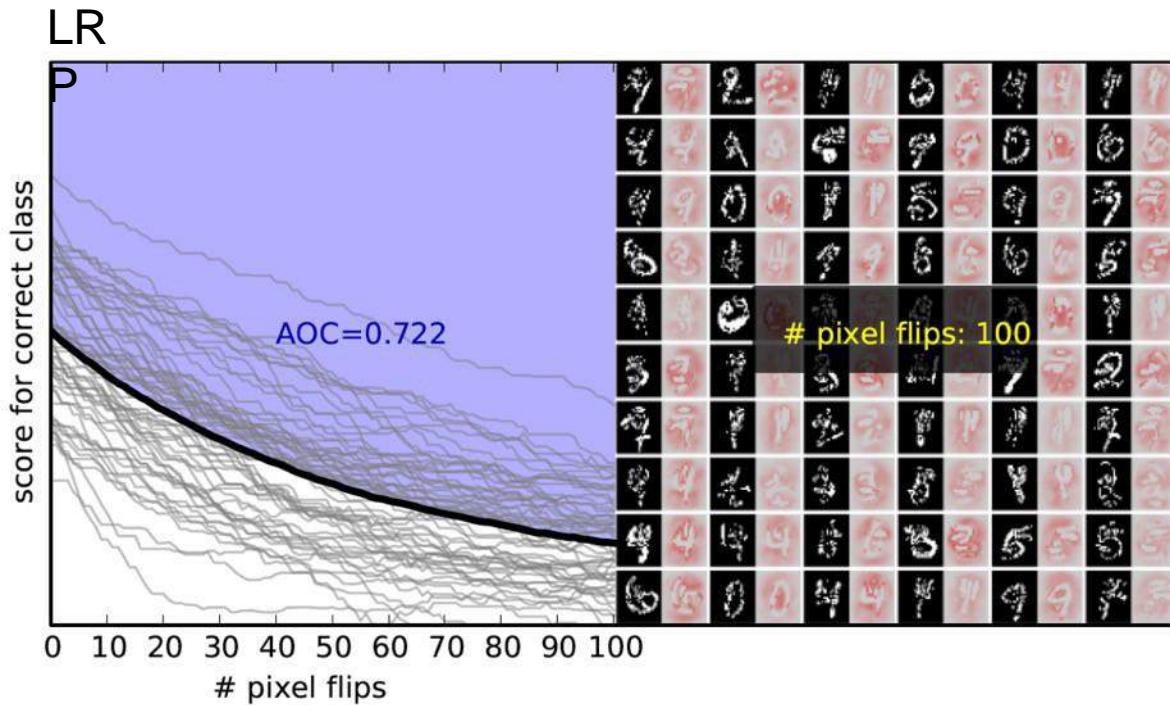
LR



Compare Explanation Methods

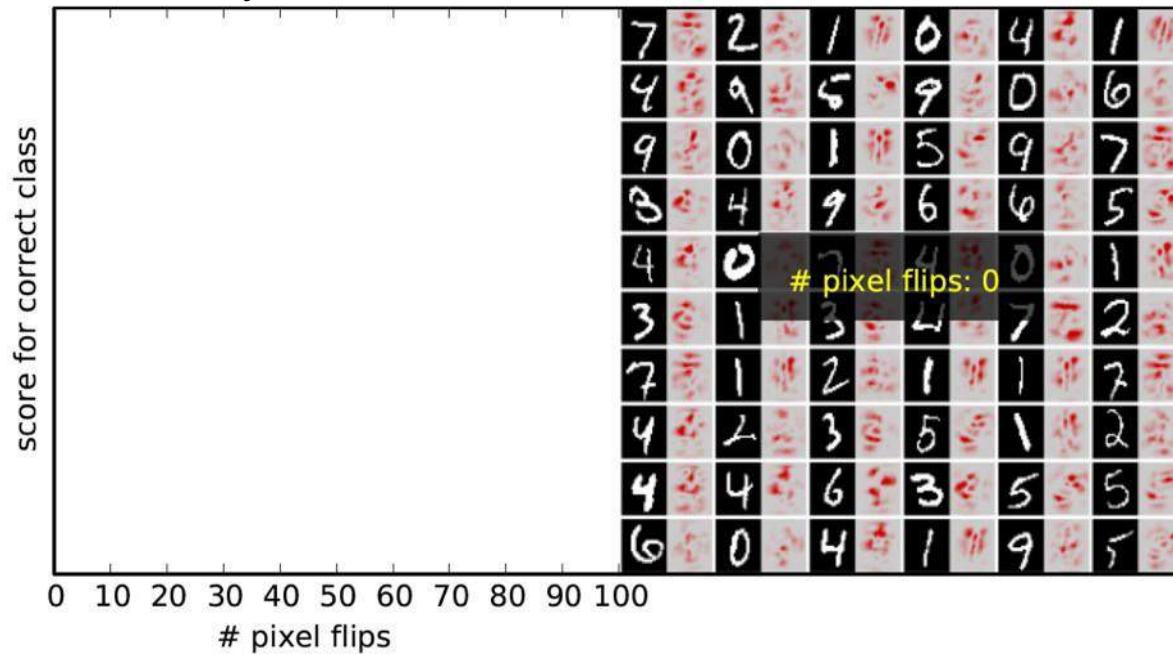


Compare Explanation Methods



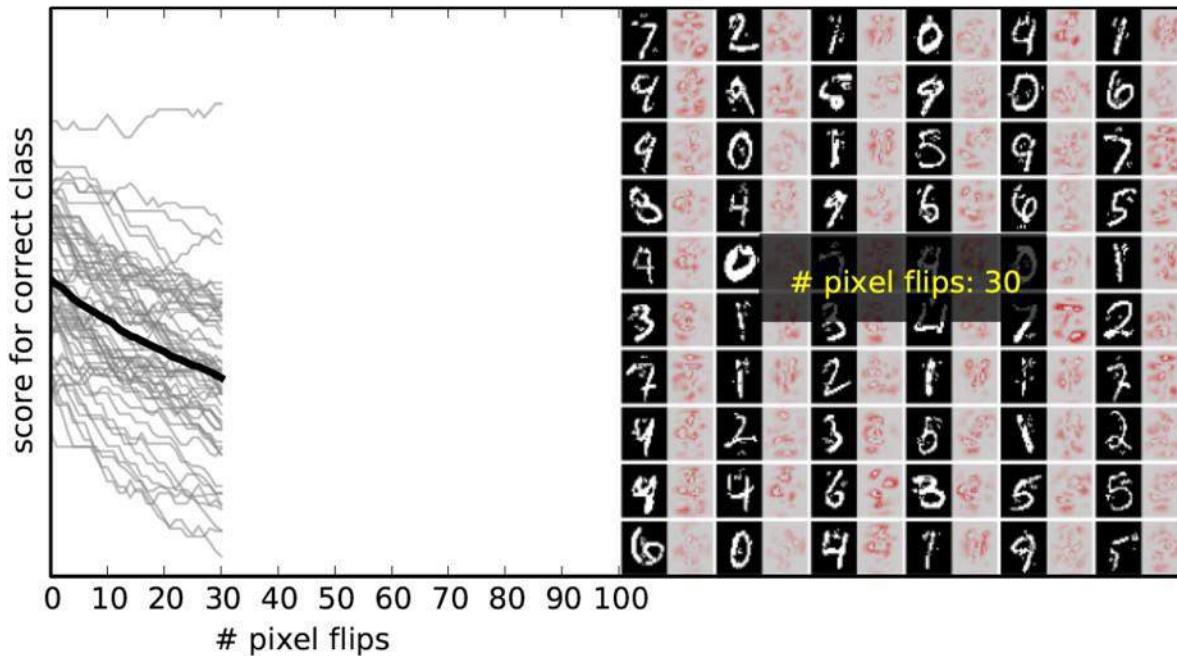
Compare Explanation Methods

Sensitivity



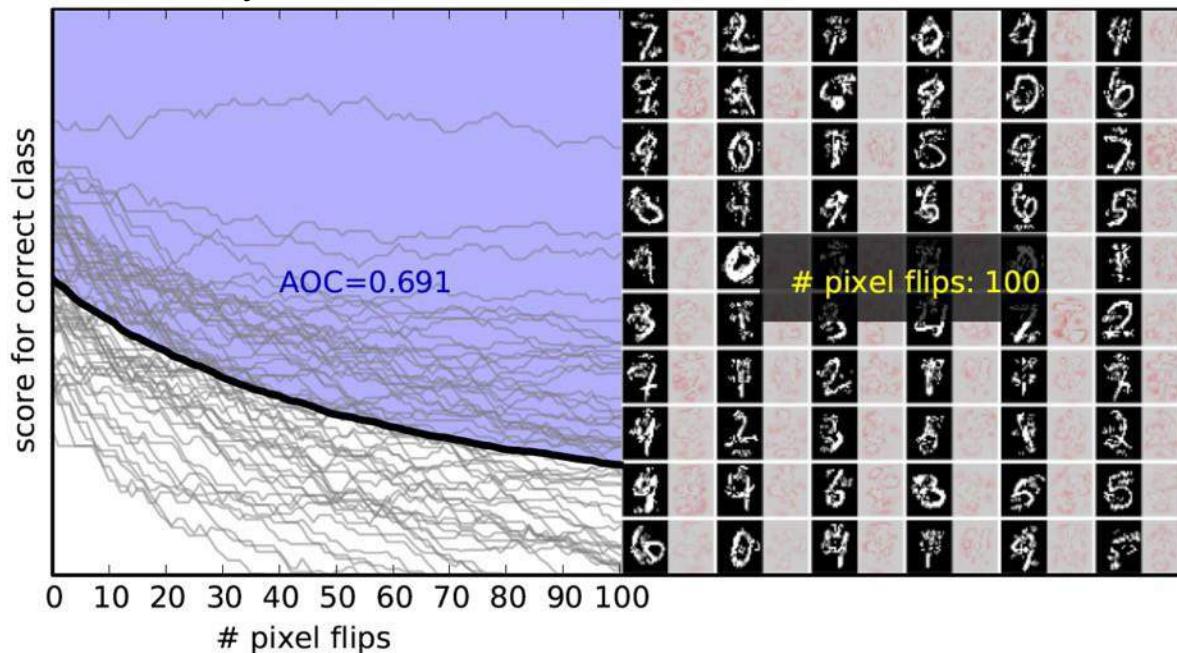
Compare Explanation Methods

Sensitivity



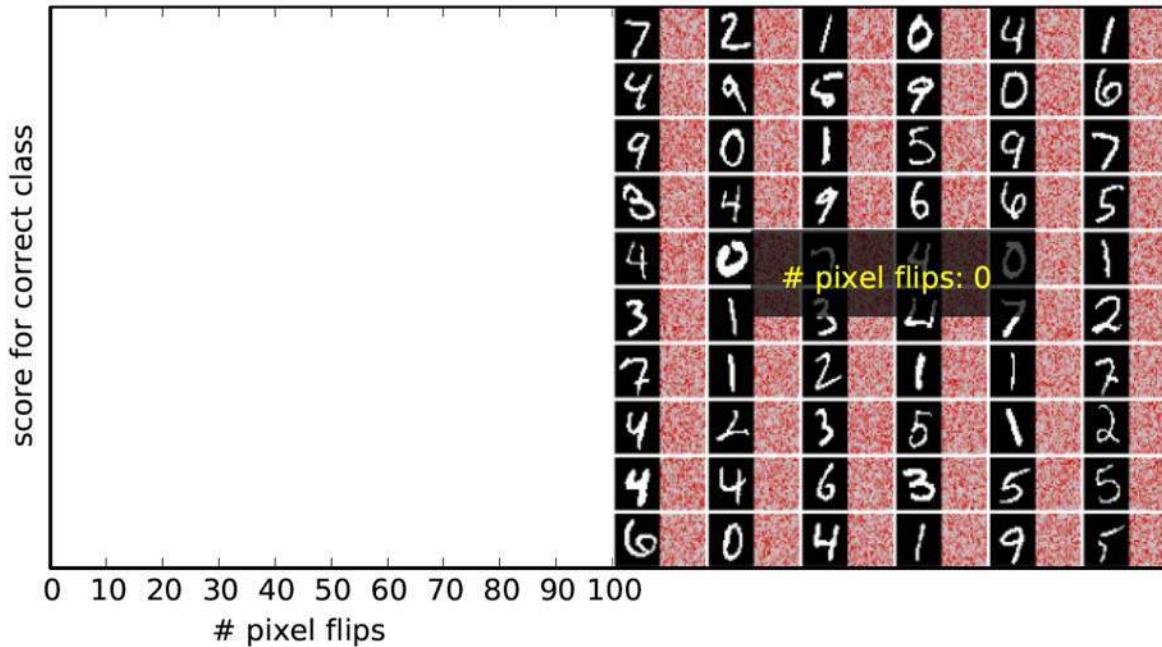
Compare Explanation Methods

Sensitivity



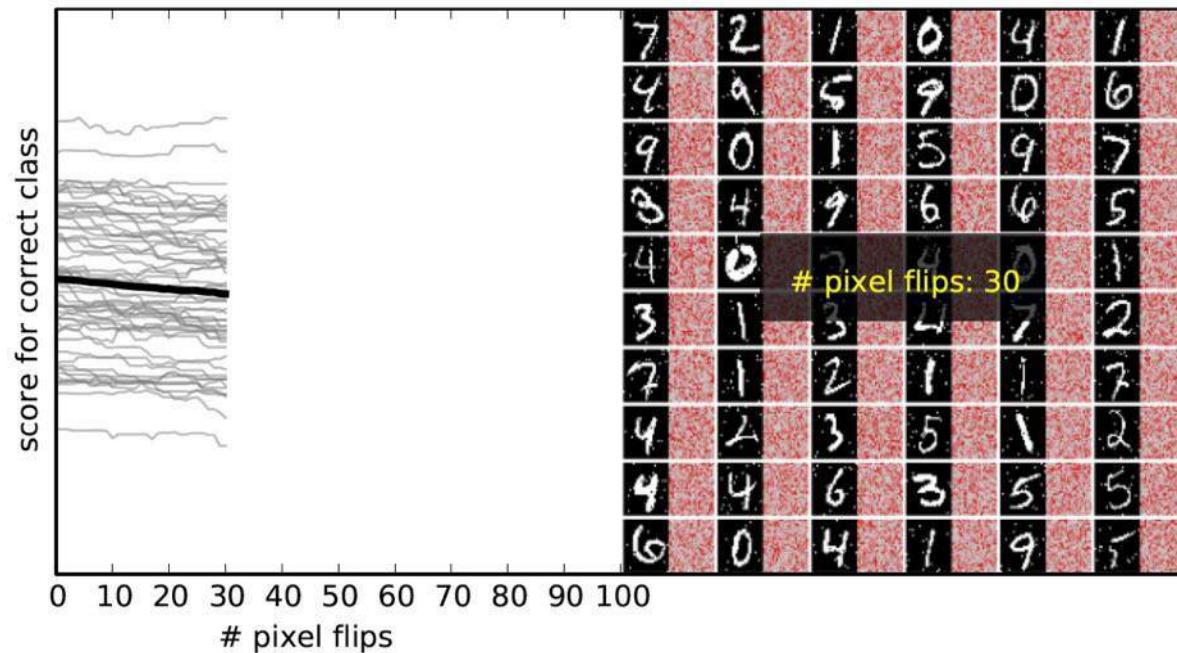
Compare Explanation Methods

Random

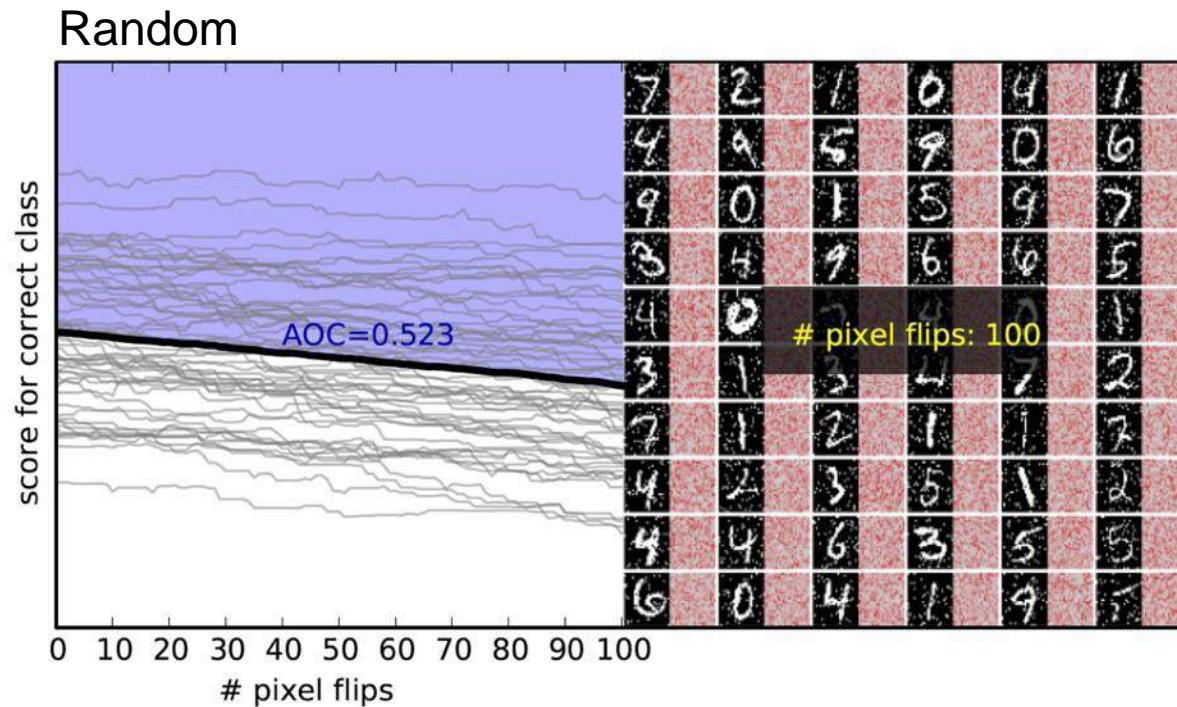


Compare Explanation Methods

Random



Compare Explanation Methods



Compare Explanation Methods

LRP:	0.722
Sensitivity:	0.691
Random:	0.523

LRP produces quantitatively better heatmaps than sensitivity analysis and random.

What about more complex datasets ?

SUN397



397 scene categories
(108,754 images in total)

ILSVRC2012



1000 categories
(1.2 million training images)

MIT Places



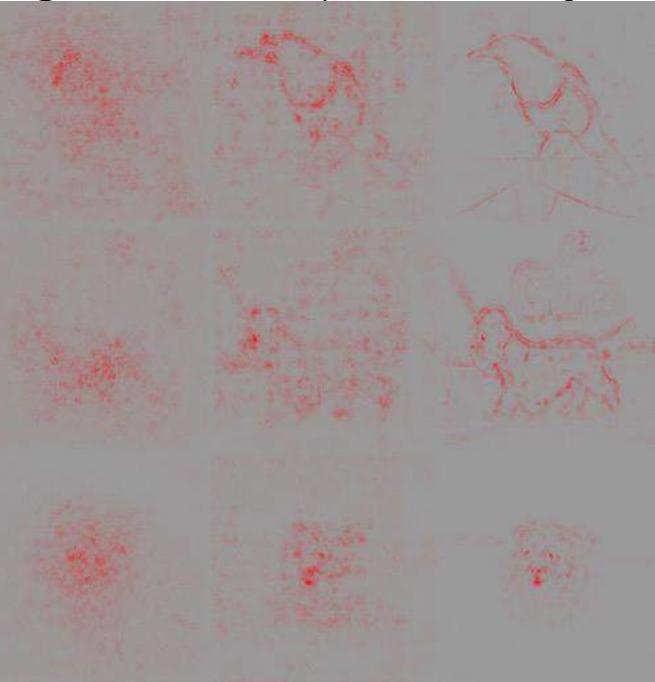
205 scene categories
(2.5 millions of images)

Compare Explanation Methods

Sensitivity Analysis
(Simonyan et al.
2014)



Deconvolution Method
(Zeiler & Fergus 2014)



LRP Algorithm
(Bach et al.
2015)

(Samek et al.
2017)

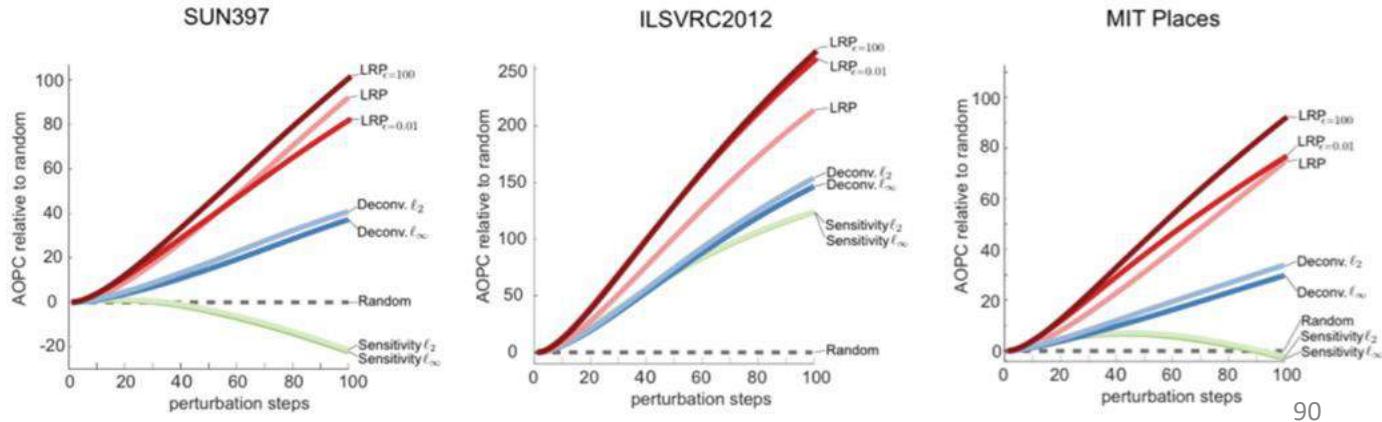
Compare Explanation Methods

Red: LRP method

Blue: Deconvolution method (Zeiler & Fergus, 2014)

Green: Sensitivity method (Simonyan et al., 2014)

- ImageNet: Caffe reference model
- Places & SUN: Classifier from MIT
- AOPC averages over 5040 images
- perturb 9×9 nonoverlapping regions
- 100 steps (15.7% of the image)
- uniform sampling in pixel space



LRP produces better heatmaps

- Sensitivity heatmaps are noisy (gradient shuttering)
- Deconvolution and sensitivity analysis solve a different problem

(Samek et al.
2017)

Compare Explanation Methods

Same idea can be applied for other domains (e.g. text document classification)

“Pixel flipping”
= “Word deleting”

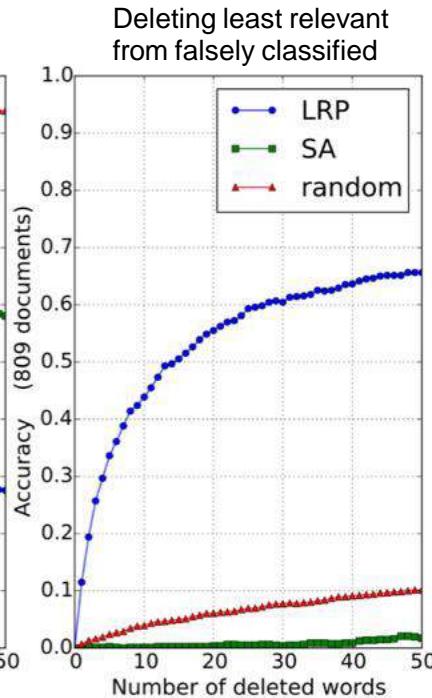
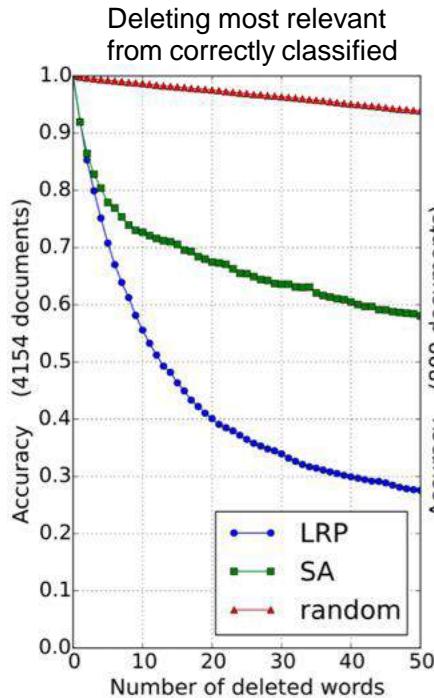
Text classified as “sci.med” —> LRP identifies most relevant words.

Yes, weightlessness does feel like falling. It may feel strange at first, but the body does adjust. The feeling is not too different from that of sky diving.

- (4.1) sci.med
- >And what is the motion sickness
>that some astronauts occasionally experience?
- It is the body's reaction to a strange environment. It appears to be induced partly to physical discomfort and part to mental distress. Some people are more prone to it than others, like some people are more prone to get sick on a roller coaster ride than others. The mental part is usually induced by a lack of clear indication of which way is up or down, ie: the Shuttle is normally oriented with its cargo bay pointed towards Earth, so the Earth (or ground) is "above" the head of the astronauts. About 50% of the astronauts experience some form of motion sickness, and NASA has done numerous tests in space to try to see how to keep the number of occurrences down.

(Arras et al.
2017)

Compare Explanation Methods



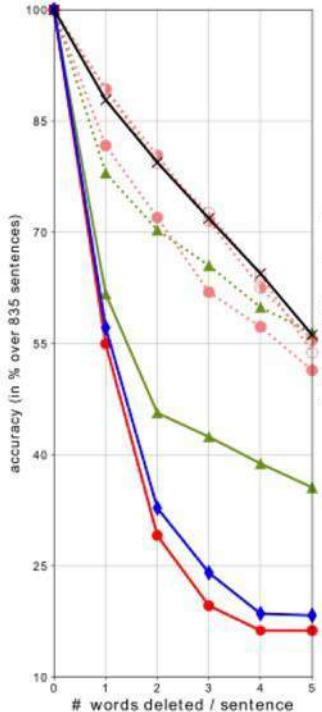
- word2vec / CNN model
- Conv → ReLU → 1-Max-Pool → FC
- trained on 20Newsgroup Dataset
- accuracy: 80.19%

LRP better than SA

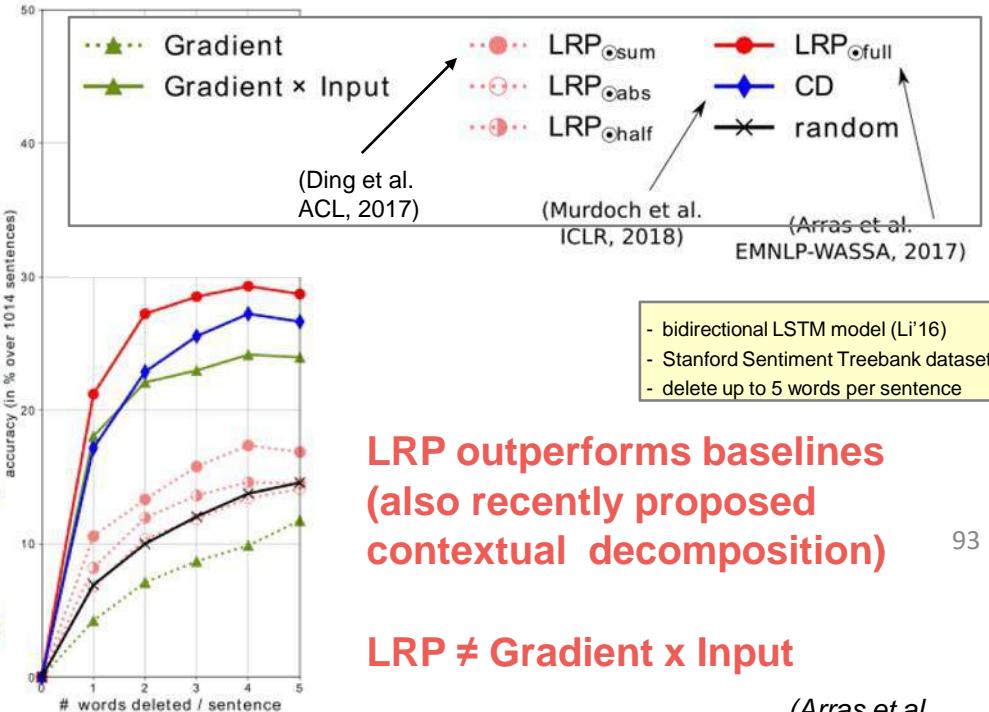
LRP distinguishes
between positive and
negative evidence

Compare Explanation Methods

Deleting most relevant
from correctly classified



Deleting least relevant
from falsely classified



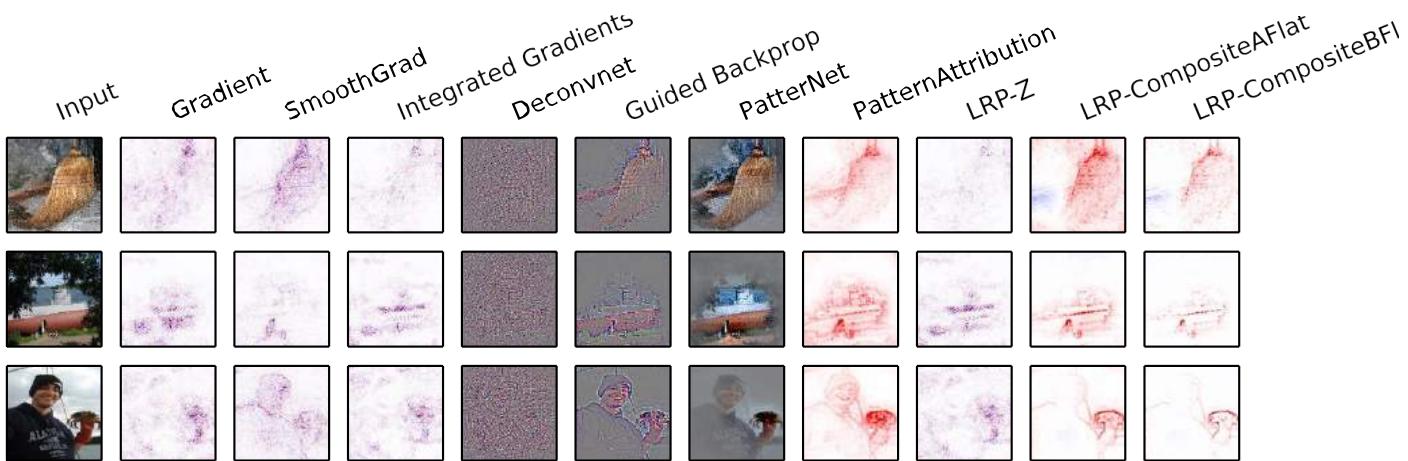
LRP outperforms baselines
(also recently proposed
contextual decomposition)

93

LRP ≠ Gradient x Input

(Arras et al.
2018)

Compare Explanation Methods



Highly efficient (e.g., 0.01 sec per VGG16 explanation)

!

New Keras Toolbox available for explanation methods:
<https://github.com/albermax/investigate>

Compare models

Application: Compare Classifiers

word2vec/CNN:

Performance: 80.19%

Strategy to solve the problem:
identify semantically meaningful words related to the topic.

BoW/SVM:

Performance: 80.10%

Strategy to solve the problem:
identify statistical patterns,
i.e., use word statistics

Yes, weightlessness does feel like falling. It may feel strange at first, but the body does adjust. The feeling is not too different from that of sky diving.

(4.1)

>And what is the motion sickness
>that some astronauts occasionally experience?

sci.med

It is the body's reaction to a strange environment. It appears to be induced partly to physical discomfort and part to mental distress. Some people are more prone to it than others, like some people are more prone to get sick on a roller coaster ride than others. The mental part is usually induced by a lack of clear indication of which way is up or down, ie: the Shuttle is normally oriented with its cargo bay pointed towards Earth, so the Earth (or ground) is "above" the head of the astronauts. About 50% of the astronauts experience some form of motion sickness, and NASA has done numerous tests in space to try to see how to keep the number of occurrences down.

(-0.6)

>And what is the motion sickness
>that some astronauts occasionally experience?

sci.med

It is the body's reaction to a strange environment. It appears to be induced partly to physical discomfort and part to mental distress. Some people are more prone to it than others, like some people are more prone to get sick on a roller coaster ride than others. The mental part is usually induced by a lack of clear indication of which way is up or down, ie: the Shuttle is normally oriented with its cargo bay pointed towards Earth, so the Earth (or ground) is "above" the head of the astronauts. About 50% of the astronauts experience some form of motion sickness, and NASA has done numerous tests in space to try to see how to keep the number of occurrences down.

96

(Arras et al. 2016 &
2017)

Application: Compare Classifiers

word2vec / CNN model

sci.med

symptoms (7.3), treatments (6.6), medication (6.4), osteopathy (6.3), ulcers (6.2), sciatica (6.0), hypertension (6.0), herb (5.6), doctor (5.4), physician (5.1), Therapy (5.1), antibiotics (5.1), Asthma (5.0), renal (5.0), medicines (4.9), caffeine (4.9), infection (4.9), gastrointestinal (4.8), therapy (4.8), homeopathic (4.7), medicine (4.7), allergic (4.7), dosages (4.7), esophagitis (4.7), inflammation (4.6), arrhythmias (4.6), cancer (4.6), disease (4.6), migraine (4.6), patients (4.5).

BoW/SVM model

sci.med

cancer (1.4), photography (1.0), doctor (1.0), msg (0.9), disease (0.9), medical (0.8), sleep (0.8), radiologist (0.7), eye (0.7), treatment (0.7), prozac (0.7), vitamin (0.7), epilepsy (0.7), health (0.6), yeast (0.6), skin (0.6), pain (0.5), liver (0.5), physician (0.5), she (0.5), needles (0.5), dn (0.5), circumcision (0.5), syndrome (0.5), migraine (0.5), antibiotic (0.5), water (0.5), blood (0.5), fat (0.4)₉₇, weight (0.4).

Words with maximum relevance

(Arras et al. 2016 & 2017)

LRP in Practice

Visual Object Classes Challenge: 2005 - 2012

	aeroplane	bicycle	bird	boat	bottle	bus	car	cat	chair	cow	diningtable	dog	horse	motorbike	person	pottedplant	sheep	sofa	train	tvmonitor	
INRIA_Flat	74.8	62.5	51.2	69.4	29.2	60.4	76.3	57.6	53.1	41.1	54.0	42.8	76.5	62.3	84.5	35.3	41.3	50.1	77.6	49.3	
INRIA_Genetic	77.5	63.6	56.1	71.9	33.1	60.6	78.0	58.8	53.5	42.6	54.9	45.8	77.5	64.0	85.9	36.3	44.7	50.6	79.2	53.2	
INRIA_Larlus	62.6	54.0	32.8	47.5	17.8	46.4	69.6	44.2	44.6	26.0	38.1	34.0	66.0	55.1	77.2	13.1	29.1	36.7	62.7	43.3	
MPI_BOW	58.9	46.0	31.3	59.0	16.9	40.5	67.2	40.2	44.3	28.3	31.9	34.4	63.6	53.5	75.7	22.3	26.6	35.4	60.6	40.6	
PRIPUVA	48.6	20.9	21.3	17.2	6.4	14.2	45.0	31.4	27.4	12.3	14.3	23.7	30.1	13.3	62.0	10.0	12.4	13.3	26.7	26.2	
QMUL_HSLS	70.6	54.8	35.7	64.5	27.8	51.1	71.4	54.0	46.6	36.6	34.4	39.9	71.5	55.4	80.6	15.8	35.8	41.5	73.1	45.5	
QMUL_LSPCH	71.6	55.0	41.1	65.5	27.2	51.1	72.2	55.1	47.4	35.9	37.4	41.5	71.5	57.9	80.8	15.6	33.3	41.9	76.5	45.9	
TKK	71.4	51.7	48.5	63.4	27.3	49.9	70.1	51.2	51.7	32.3	46.3	41.5	72.6	60.2	82.2	31.7	30.1	39.2	71.1	41.0	
ToshCam_rdf	59.9	36.8	29.9	40.0	23.6	33.3	60.2	33.0	41.0	17.8	33.2	33.7	63.9	53.1	77.9	29.0	27.3	31.2	50.1	37.6	
ToshCam_svm	54.0	27.1	30.3	35.6	17.0	22.3	58.0	34.6	38.0	19.0	27.5	32.4	48.0	40.7	78.1	23.4	21.8	28.0	45.5	31.8	
Tsinghua	62.9	42.4	33.9	49.7	23.7	40.7	62.0	35.2	42.7	21.0	38.9	34.7	65.0	48.1	76.9	16.9	30.8	32.8	58.9	33.1	
UVA_Bigrams	61.2	33.2	29.4	45.0	16.5	37.6	54.6	31.3	39.9	17.2	31.4	30.6	61.6	42.4	74.6	14.5	20.9	23.5	49.9	30.0	
UVA_FuseAll	67.1	48.1	43.3	58.1	19.9	46.3	61.8	41.9	48.4	27.8	41.9	38.5	69.8	51.4	79.4	32.5	31.9	36.0	66.2	40.3	
UVA_MCIP	66.5	47.9	41.0	58.0	16.8	44.0	61.2	40.5	48.5	27.8	41.7	37.1	66.4	50.1	78.6	31.2	32.3	31.9	66.6	40.3	
UVA_SFS	66.3	49.7	43.5	60.7	18.8	44.9	64.8	41.9	46.8	24.9	42.3	33.9	71.5	53.4	80.4	29.7	31.2	31.8	67.4	43.5	
UVA_WGT	59.7	33.7	34.9	44.5	22.2	32.9	55.9	36.3	36.8	20.6	25.2	34.7	65.1	40.1	74.2	26.4	26.9	25.1	50.7	29.7	
XRCE	72.3	57.5	53.2	68.9	28.5	57.5	75.4	50.3	52.2	39.0	46.8	45.3	75.7	58.5	84.0	32.6	39.7	50.9	75.1	49.5	98

Application: Compare Classifiers

Test error for various classes:

	aeroplane	bicycle	bird	boat	bottle	bus	car
Fisher	79.08%	66.44%	45.90%	70.88%	27.64%	69.67%	80.96%
DeepNet	88.08%	79.69%	80.77%	77.20%	35.48%	72.71%	86.30%
	cat	chair	cow	diningtable	dog	horse	motorbike
Fisher	59.92%	51.92%	47.60%	58.06%	42.28%	80.45%	69.34%
DeepNet	81.10%	51.04%	61.10%	64.62%	76.17%	81.60%	79.33%
	person	pottedplant	sheep	sofa	train	tvmonitor	mAP
Fisher	85.10%	28.62%	49.58%	49.31%	82.71%	54.33%	59.99%
DeepNet	92.43%	49.99%	74.04%	49.48%	87.07%	67.08%	72.12%

(Lapuschkin et al.
2016)

Application: Compare Classifiers

Test error for various classes:

	aeroplane	bicycle	bird	boat	bottle	bus	car
Fisher	79.08%	66.44%	45.90%	70.88%	27.64%	69.67%	80.96%
DeepNet	88.08%	79.69%	80.77%	77.20%	35.48%	72.71%	86.30%
	cat	chair	cow	diningtable	dog	horse	motorbike
Fisher	59.92%	51.92%	47.60%	58.06%	42.28%	80.45%	69.34%
DeepNet	81.10%	51.04%	61.10%	64.62%	76.17%	81.60%	79.33%
	person	pottedplant	sheep	sofa	train	tvmonitor	mAP
Fisher	85.10%	28.62%	49.58%	49.31%	82.71%	54.33%	59.99%
DeepNet	92.43%	49.99%	74.04%	49.48%	87.07%	67.08%	72.12%

same performance → same strategy ?

(Lapuschkin et al.
2016)

Application: Compare Classifiers

Test error for various classes:

	aeroplane	bicycle	bird	boat	bottle	bus	car
Fisher	79.08%	66.44%	45.90%	70.88%	27.64%	69.67%	80.96%
DeepNet	88.08%	79.69%	80.77%	77.20%	35.48%	72.71%	86.30%
	cat	chair	cow	diningtable	dog	horse	motorbike
Fisher	59.92%	51.92%	47.60%	58.06%	42.28%	80.45%	69.34%
DeepNet	81.10%	51.04%	61.10%	64.62%	76.17%	81.60%	79.33%
	person	pottedplant	sheep	sofa	train	tvmonitor	mAP
Fisher	85.10%	28.62%	49.58%	49.31%	82.71%	54.33%	59.99%
DeepNet	92.43%	49.99%	74.04%	49.48%	87.07%	67.08%	72.12%

Image



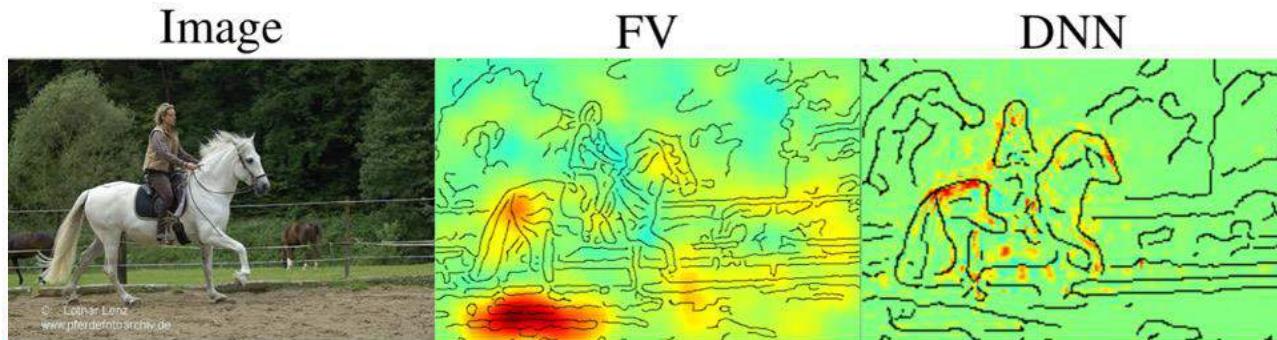
same performance → same strategy ?

(Lapuschkin et al.
2016)

Application: Compare Classifiers

Test error for various classes:

	aeroplane	bicycle	bird	boat	bottle	bus	car
Fisher	79.08%	66.44%	45.90%	70.88%	27.64%	69.67%	80.96%
DeepNet	88.08%	79.69%	80.77%	77.20%	35.48%	72.71%	86.30%
	cat	chair	cow	diningtable	dog	horse	motorbike
Fisher	59.92%	51.92%	47.60%	58.06%	42.28%	80.45%	69.34%
DeepNet	81.10%	51.04%	61.10%	64.62%	76.17%	81.60%	79.33%
	person	pottedplant	sheep	sofa	train	tvmonitor	mAP
Fisher	85.10%	28.62%	49.58%	49.31%	82.71%	54.33%	59.99%
DeepNet	92.43%	49.99%	74.04%	49.48%	87.07%	67.08%	72.12%



same performance → same strategy ?

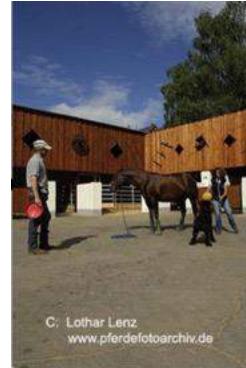
(Lapuschkin et al.
2016)

Application: Compare Classifiers

'horse' images in PASCAL VOC 2007



C: Lothar Lenz
www.pferdefotoarchiv.de



C: Lothar Lenz
www.pferdefotoarchiv.de



C: Lothar Lenz
www.pferdefotoarchiv.de

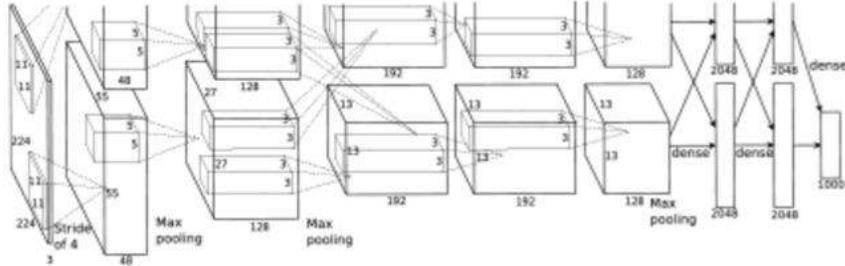


C: Lothar Lenz
www.pferdefotoarchiv.de



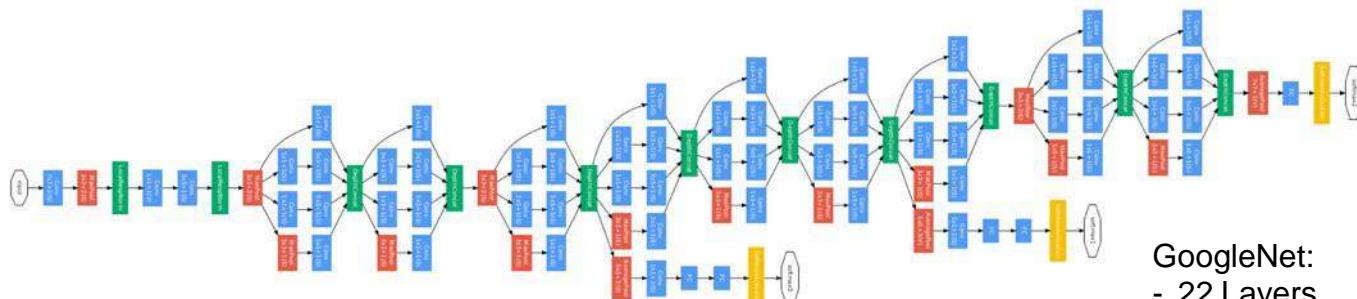
C: Lothar Lenz
www.pferdefotoarchiv.de

Application: Compare Classifiers



BVLC:

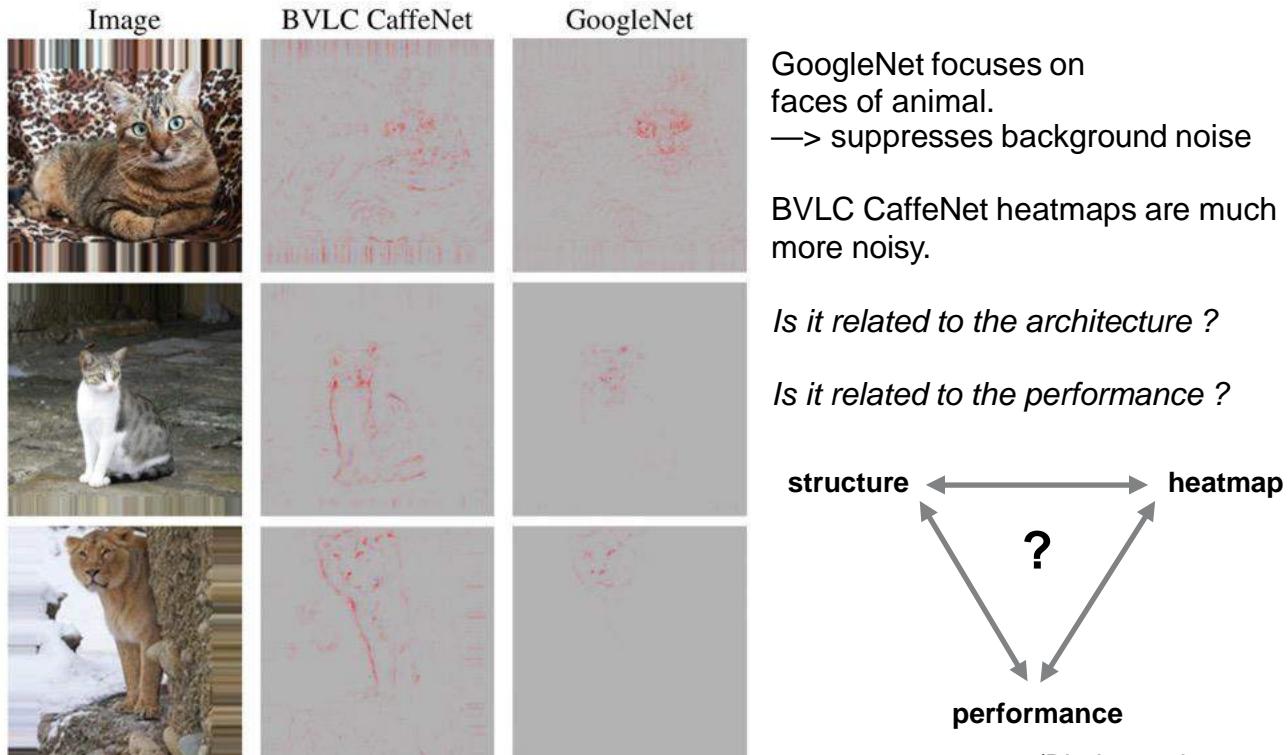
- 8 Layers
- ILSRCV: 16.4%



GoogleNet:

- 22 Layers
- ILSRCV: 6.7%
- Inception layers

Application: Compare Classifiers



(Binder et al.
2016)

Quantify Context Use

Application: Measure Context Use



how important
is context ?

classifier

how important
is context ?

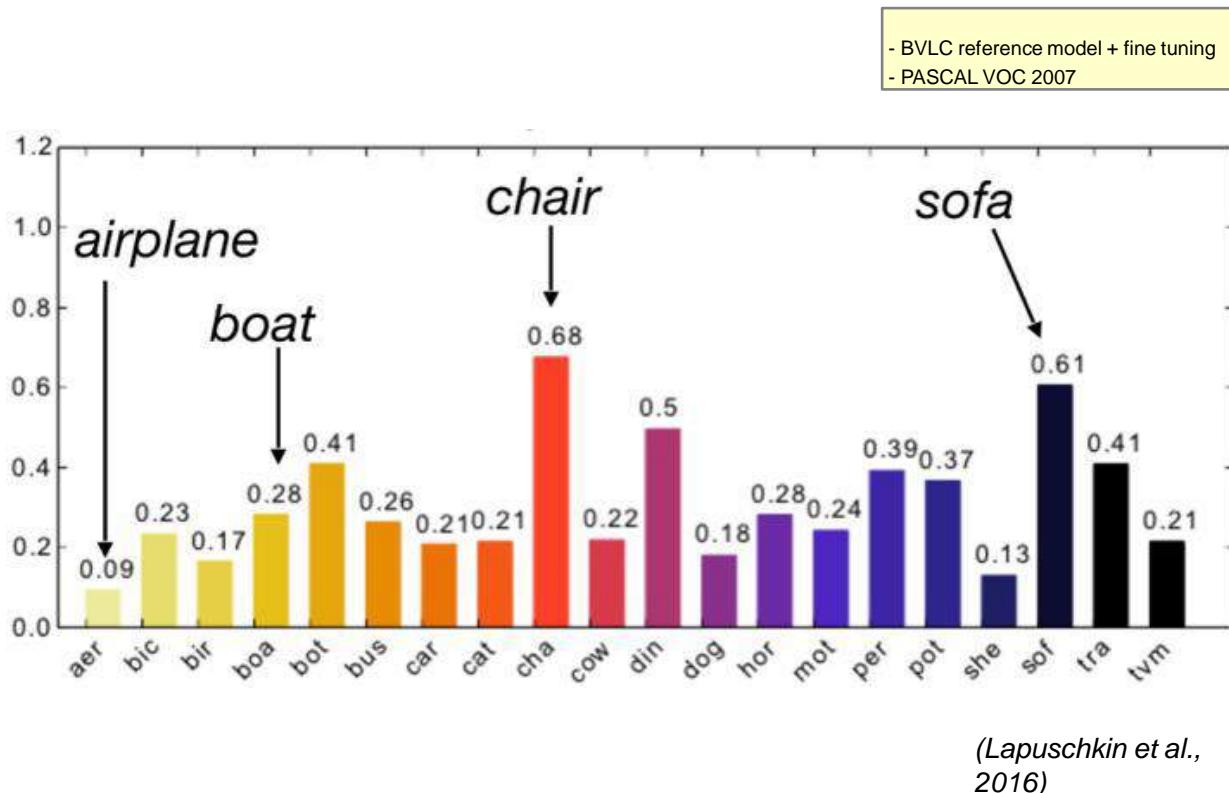
LRP decomposition allows
meaningful pooling over bbox

!

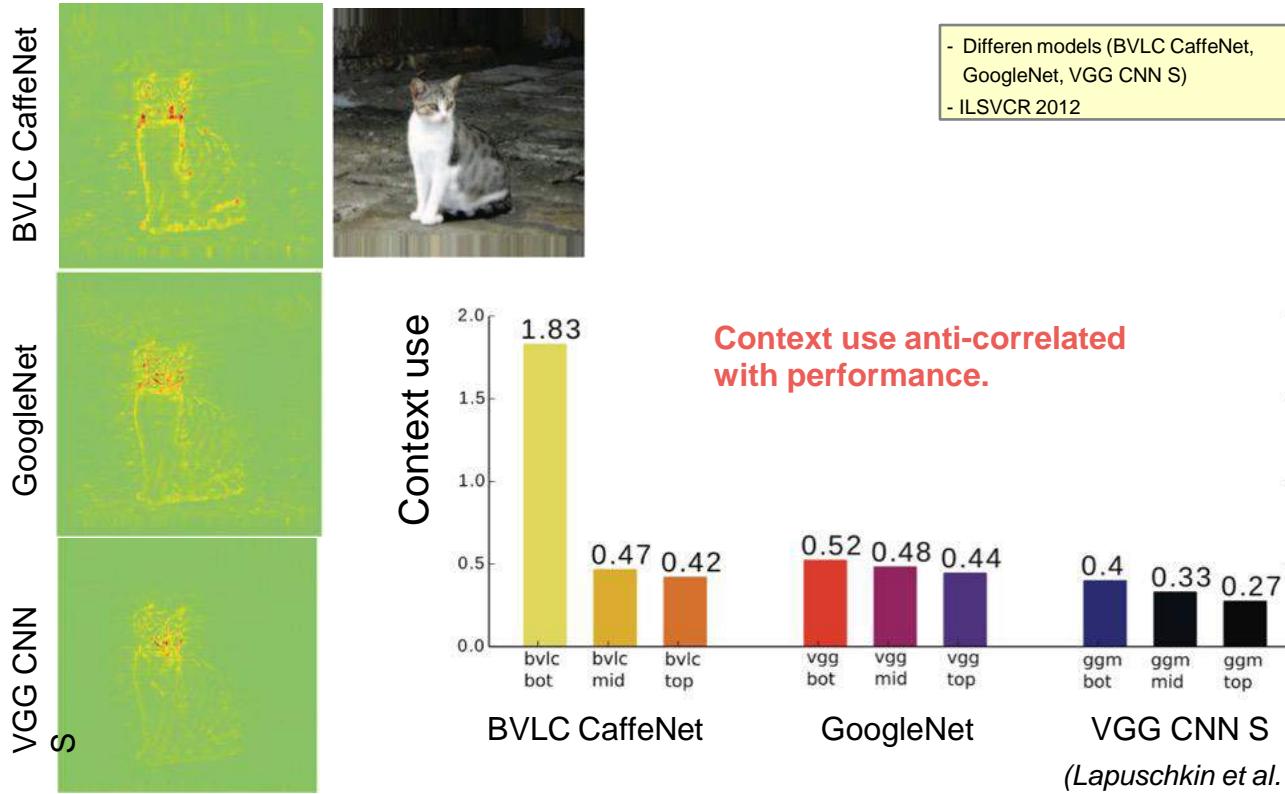
$$\sum_i R_i = f(x)$$

$$\text{importance of context} = \frac{\text{relevance outside bbox}}{\text{relevance inside bbox}}$$

Application: Measure Context Use



Application: Measure Context Use



Detect Biases & Improve Models

Application: Face analysis

- Compare AdienceNet, CaffeNet, GoogleNet, VGG-16
- Adience dataset, 26,580 images

Age classification

	A	C	G	V
[i]	51.4 87.0	52.1 87.9	54.3 89.1	–
[r]	51.9 87.4	52.3 88.9	53.3 89.9	–
[m]	53.6 88.4	54.3 89.7	56.2 90.7	–
[i,n]	–	51.6 87.4	56.2 90.9	53.6 88.2
[r,n]	–	52.1 87.0	57.4 91.9	–
[m,n]	–	52.8 88.3	58.5 92.6	56.5 90.0
[i,w]	–	–	–	59.7 94.2
[r,w]	–	–	–	–
[m,w]	–	–	–	62.8 95.8

Gender classification

	A	C	G	V
[i]	88.1	87.4	87.9	–
[r]	88.3	87.8	88.9	–
[m]	89.0	88.8	89.7	–
[i,n]	–	89.9	91.0	92.0
[r,n]	–	90.6	91.6	–
[m,n]	–	90.6	91.7	92.6
[i,w]	–	–	–	90.5
[r,w]	–	–	–	–
[m,w]	–	–	–	92.2

A = AdienceNet
 C = CaffeNet
 G = GoogleNet
 V = VGG-16

[i] = in-place face alignment
 [r] = rotation based alignment
 [m] = mixing aligned images for training
 [n] = initialization on Imagenet
 [w] = initialization on IMDB-WIKI

(Lapuschkin et al., 2017)

Application: Face analysis

Gender classification



Strategy to solve the problem: Focus on chin / beard, eyes & hair,
but without pretraining the model overfits

(Lapuschkin et al.,
2017)

Application: Face analysis

Age classification



Predictions

25-32 years old

Strategy to solve the problem:
Focus on the laughing ...



60+ years old

laughing speaks against 60+
(i.e., model learned that old
people do not laugh)

*(Lapuschkin et al.,
2017)*

Application: Face analysis

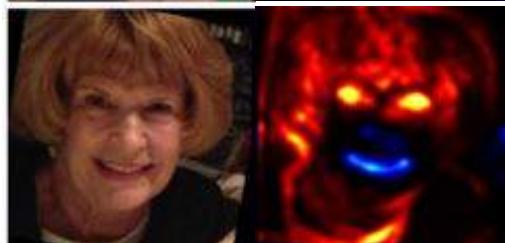
Age classification



Predictions

25-32 years old

Strategy to solve the problem:
Focus on the laughing ...



60+ years old
pretraining on

ImageNet

laughing speaks against 60+
(i.e., model learned that old
people do not laugh)

pretraining on
IMDB-WIKI



*(Lapuschkin et al.,
2017)*

Application: Face analysis



- 1,900 images of different individuals
- pretrained VGG19 model
- different ways to train the models

Different training methods

	naive	one morphed	complex morphs	multiclass
true positive	95%	90%	93%	92%
true negative	98%	95%	95%	99%
EER	3.1%	7.2%	6.1%	2.8%

50% genuine images,
50% complete morphs

50% genuine images,
10% complete morphs and
4 × 10% one region morphed

50% genuine images,
10% complete morphs,
partial morphs with 10%
one, two, three and four
region morphed

partial morphs with zero,
one, two, three or four
morphed regions,
for two class classification
last layer reinitialized

(Seibold et al.,
2018)

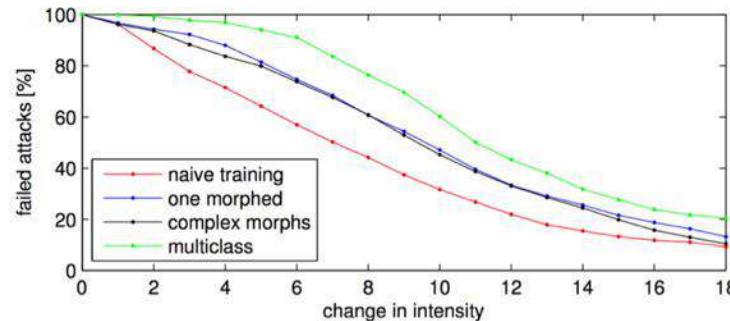
Application: Face analysis

Semantic attack
on the model

Table 4. Robustness against partial morphs.

	left eye	right eye	nose	mouth	average
naive	25%	21%	14%	13%	20%
one morphed	81%	89%	79%	71%	80%
complex morphs	78%	74%	73%	54%	70%
multiclass	86%	93%	90%	79%	87%

Black box adversarial
attack on the model



112

Fig. 5. Robustness against fast gradient sign attacks.

Application: Face analysis

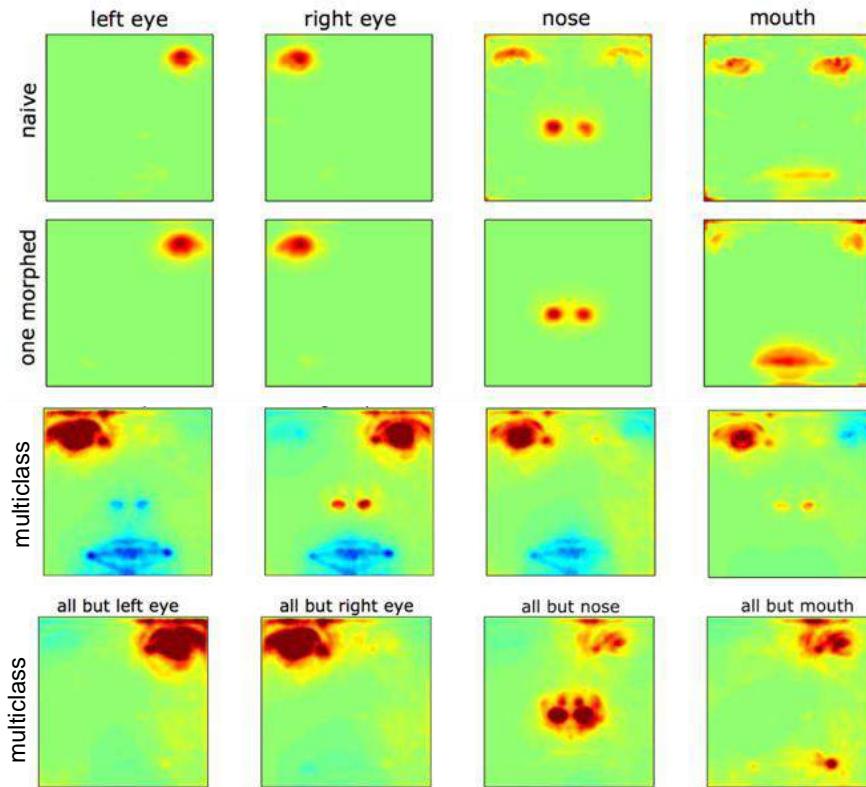
morphed region	relative amount of relevance per region							
	naive				one morphed			
	left eye	right eye	nose	mouth	left eye	right eye	nose	mouth
left eye	0.84	0.00	0.02	0.14	0.96	0.00	0.01	0.04
right eye	0.00	0.91	0.05	0.05	0.00	0.92	0.01	0.07
nose	0.21	0.28	0.47	0.04	0.00	0.01	0.97	0.02
mouth	0.34	0.27	0.04	0.35	0.17	0.12	0.04	0.68

	complex morphs				multiclass			
	left eye	right eye	nose	mouth	left eye	right eye	nose	mouth
	left eye	right eye	nose	mouth	left eye	right eye	nose	mouth
left eye	0.98	0.00	0.00	0.02	0.00	0.98	0.00	0.01
right eye	0.00	0.92	0.00	0.08	0.98	0.00	0.02	0.00
nose	0.02	0.03	0.92	0.02	0.01	0.10	0.19	0.70
mouth	0.06	0.00	0.41	0.53	0.11	0.18	0.58	0.13

113

(Seibold et al.,
2018)

Application: Face analysis



Different models
have different
strategies !

network seems to
compare different
structures

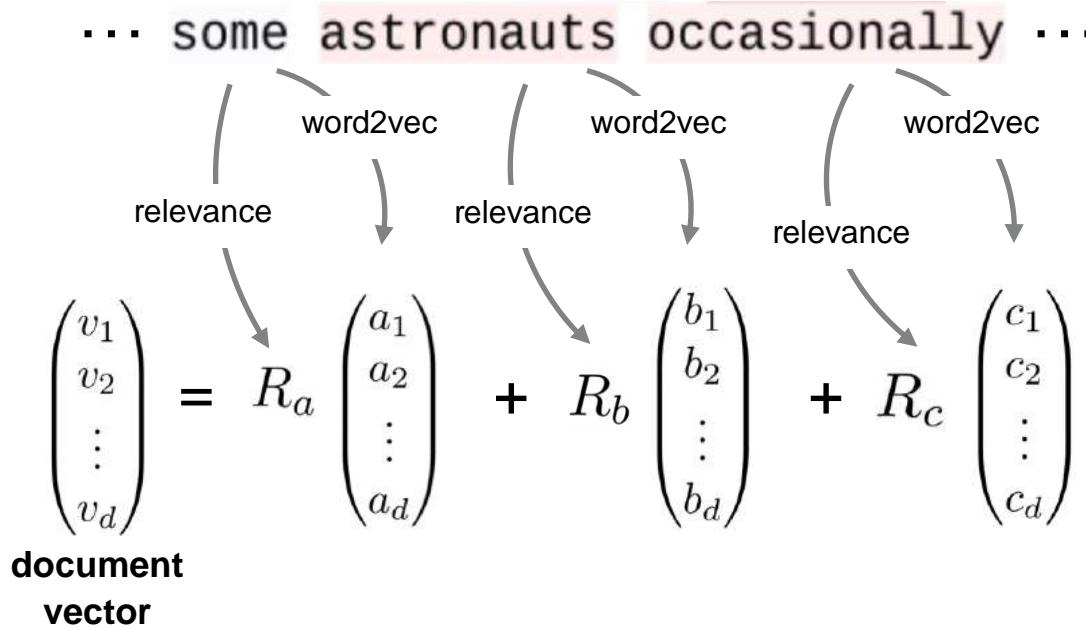
114

network seems to
identify “original”
parts

(Seibold et al.,
2018)

Learn new Representations

Application: Learn new Representations

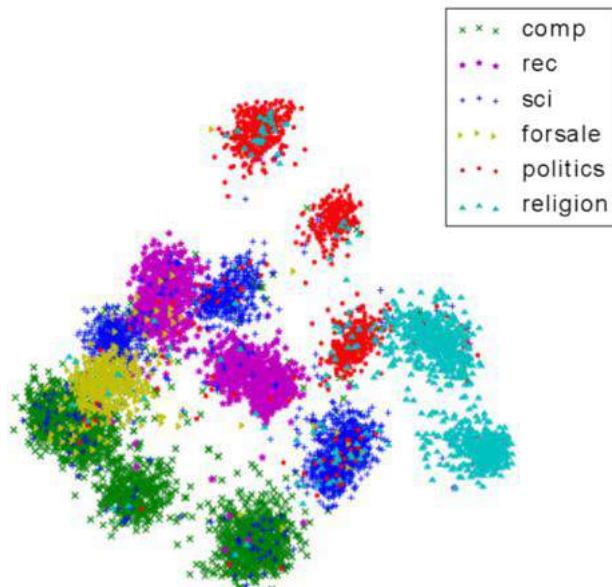


116

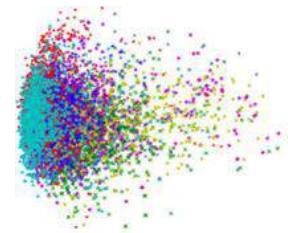
(Arras et al. 2016 & 2017)

Application: Learn new Representations

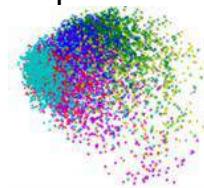
2D PCA projection of document vectors



uniform



TFID
F



Document vector computation is unsupervised (given we have a classifier).

117

(Arras et al. 2016 & 2017)

Interpreting Scientific Data

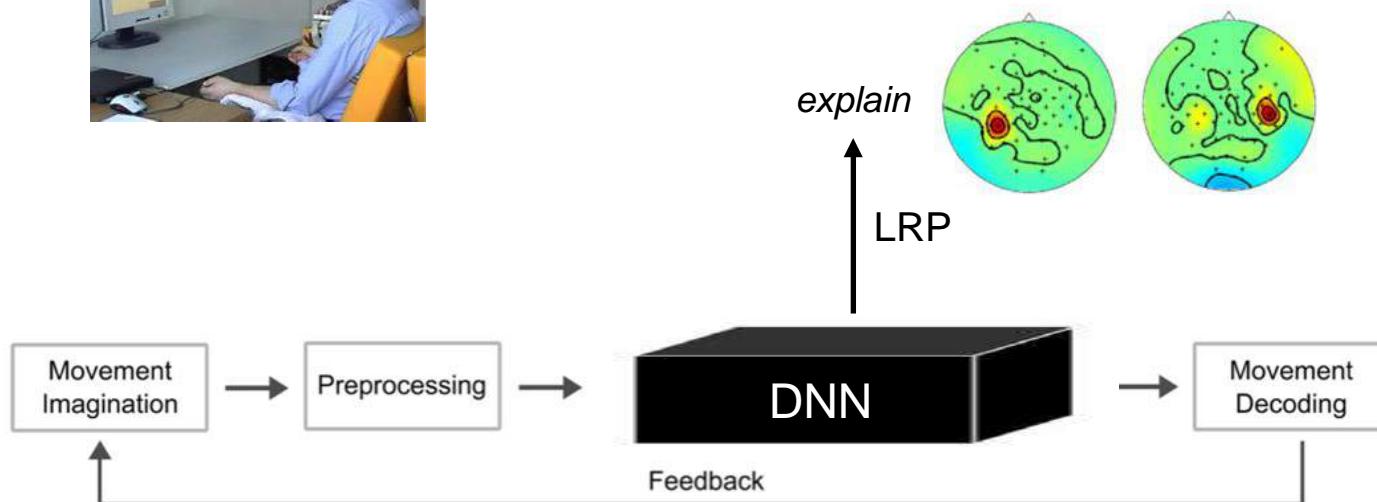
Application: EEG Analysis

Brain-Computer Interfacing



Neural network learns that:

Left hand movement imagination leads to desynchronization over right sensorimotor cortex (and vice versa).

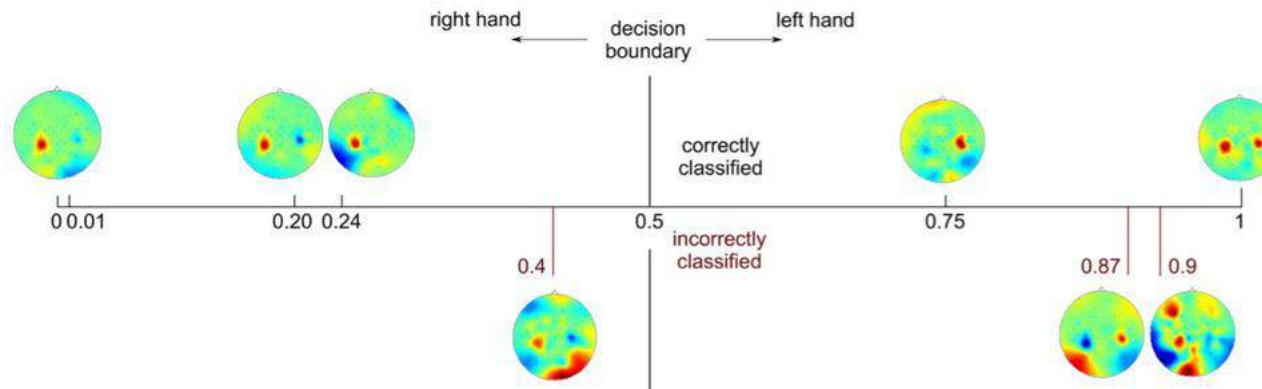


(Sturm et al.
2016)

Application: EEG Analysis

Our neural networks are interpretable:

We can see for every trial “why” it is classified the way it is.



(Sturm et al.
2016)

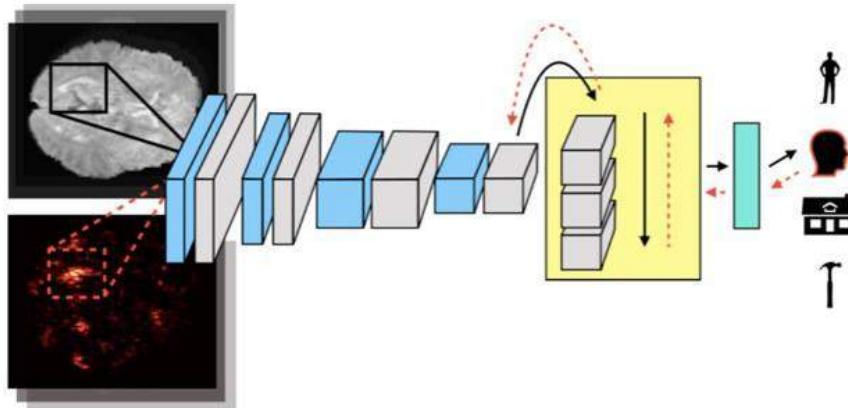
Application: fMRI Analysis

Difficulty to apply deep learning to fMRI :

- high dimensional data (100 000 voxels), but only few subjects
- results must be interpretable (key in neuroscience)

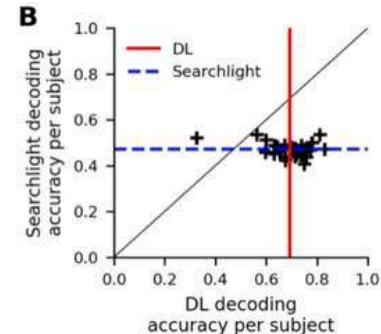
Our approach:

- Recurrent neural networks (CNN + LSTM) for whole-brain analysis
- LRP allows to interpret the results

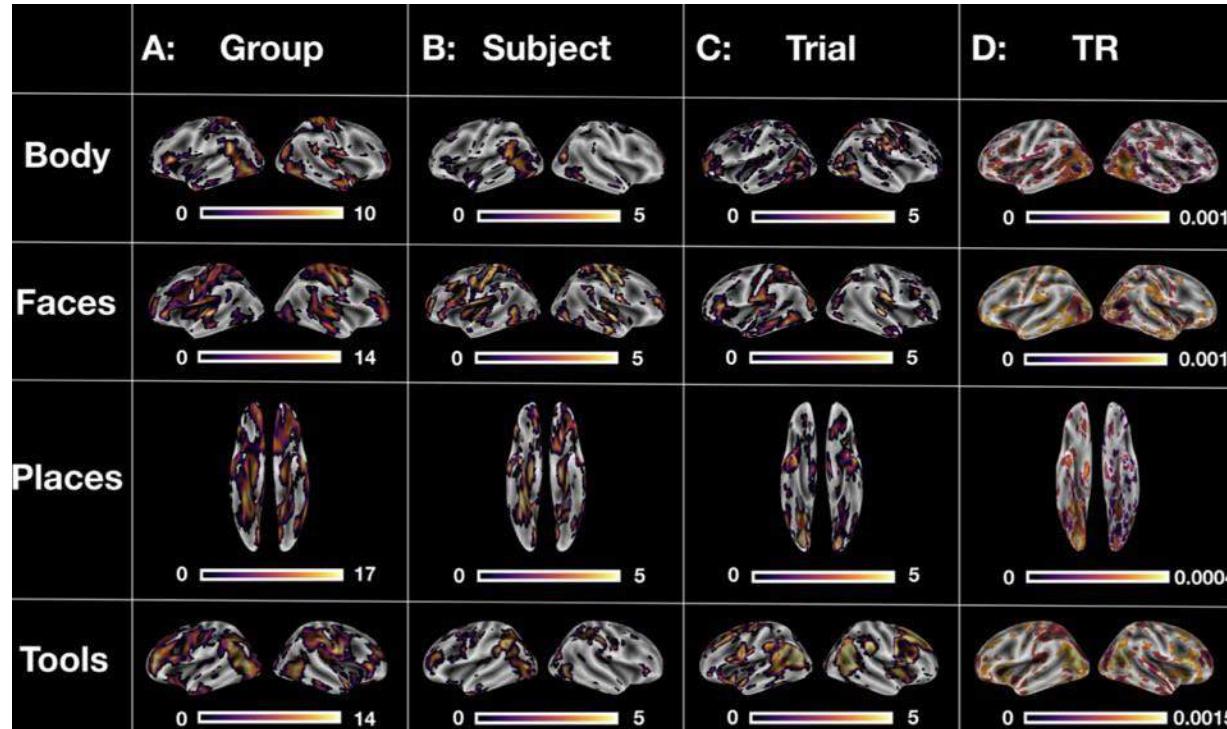


Dataset:

- 100 subjects from Human Connectome Project
- N-back task (faces, places, tools and body parts)

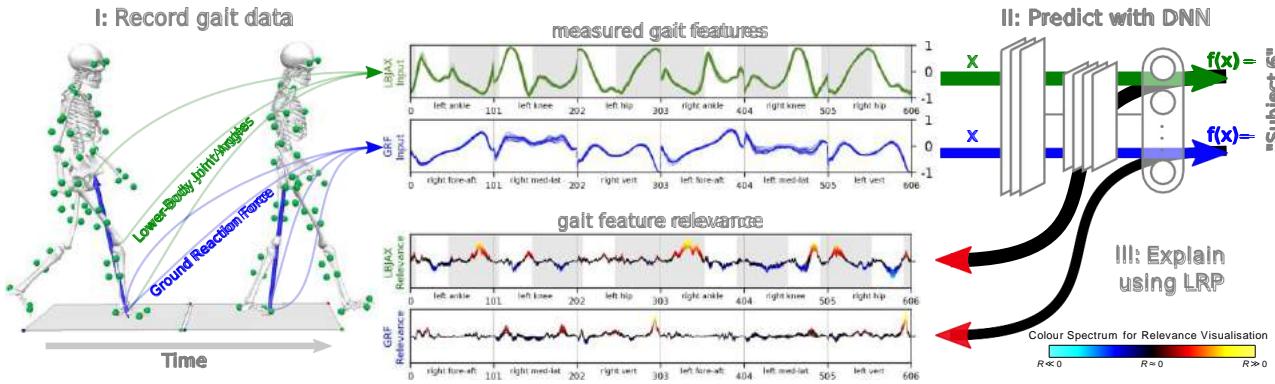


Application: fMRI Analysis



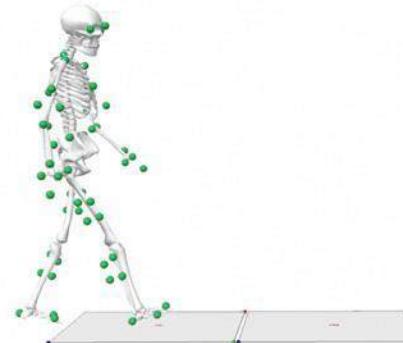
(Thomas et al.
2018)

Application: Gait Analysis



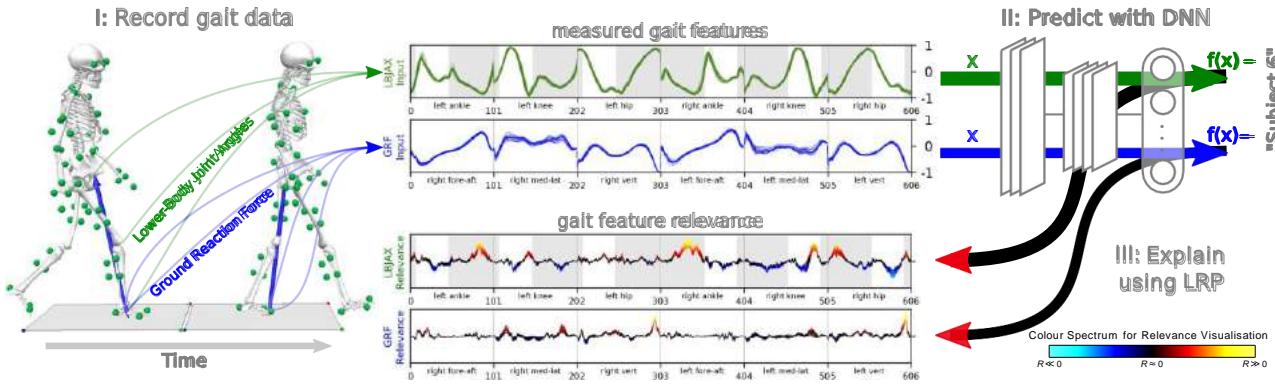
Our approach:

- Classify & explain individual gait patterns
- Important for understanding diseases such as Parkinson



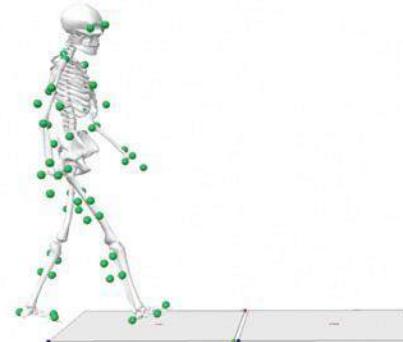
(Horst et al.
2018)

Application: Gait Analysis



Our approach:

- Classify & explain individual gait patterns
- Important for understanding diseases such as Parkinson

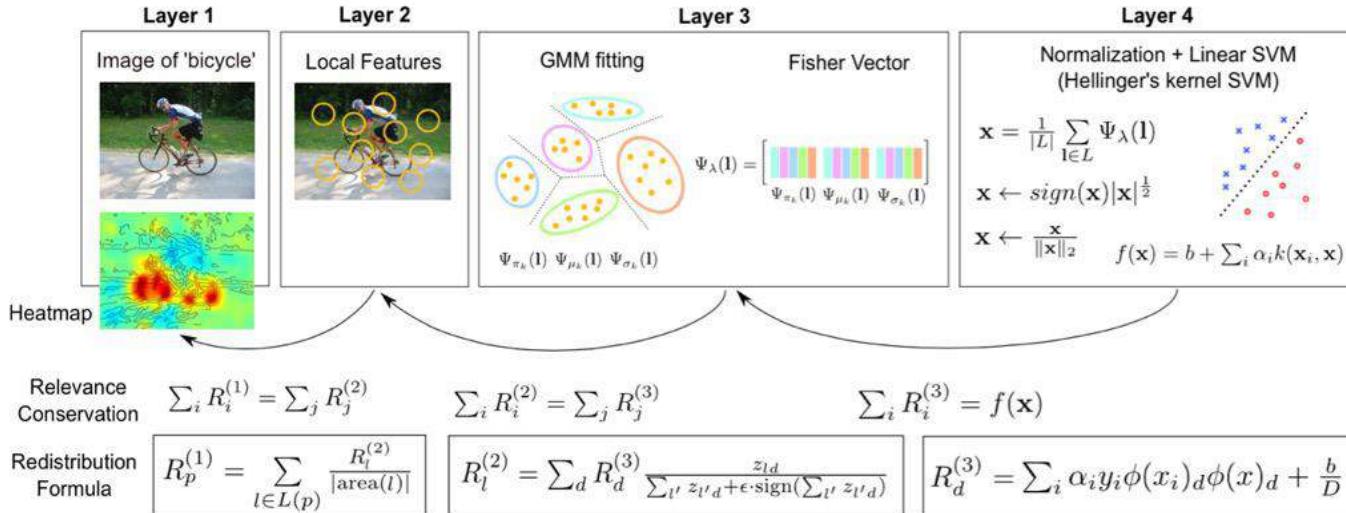


(Horst et al.
2018)

Understand Model & Obtain new Insights

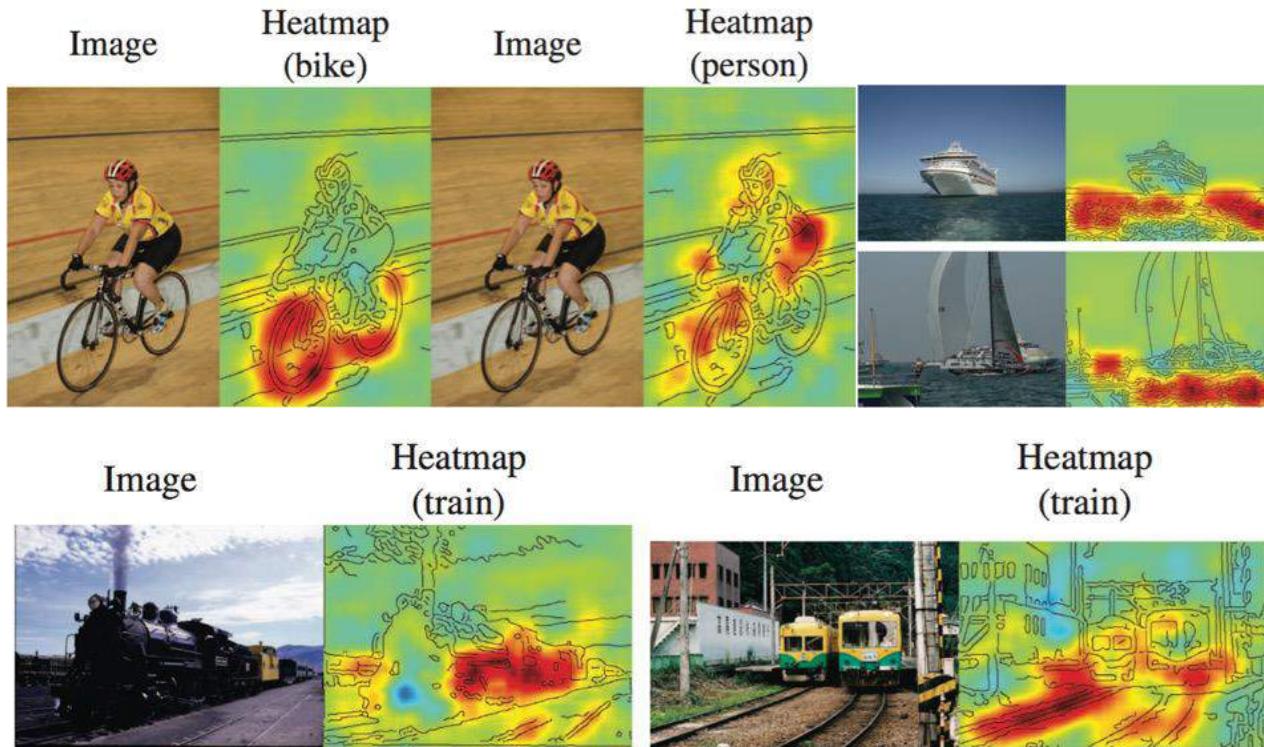
Application: Understand the model

- Fisher Vector / SVM classifier
- PASCAL VOC 2007



(Lapuschkin et al.
2016)

Application: Understand the model



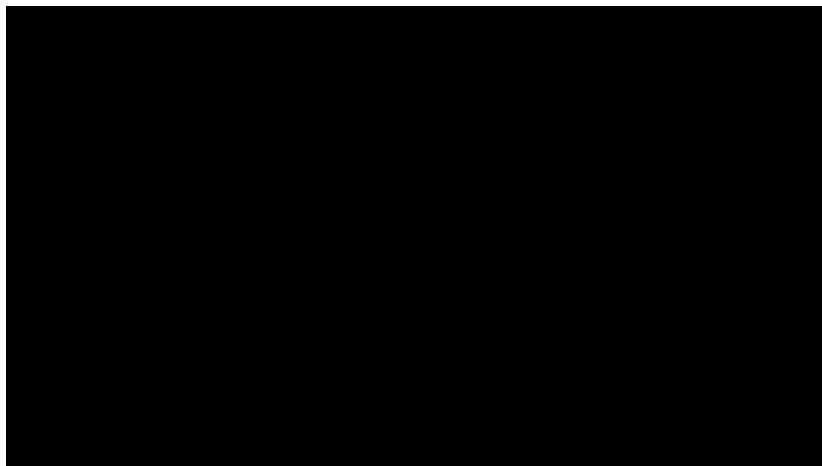
(Lapuschkin et al.
2016)

Application: Understand the model



Motion vectors can be extracted
from the compressed video
-> allows very efficient analysis

- Fisher Vector / SVM classifier
- Model of Kantorov & Laptev, (CVPR'14)
- Histogram Of Flow, Motion Boundary Histogram
- HMDB51 dataset



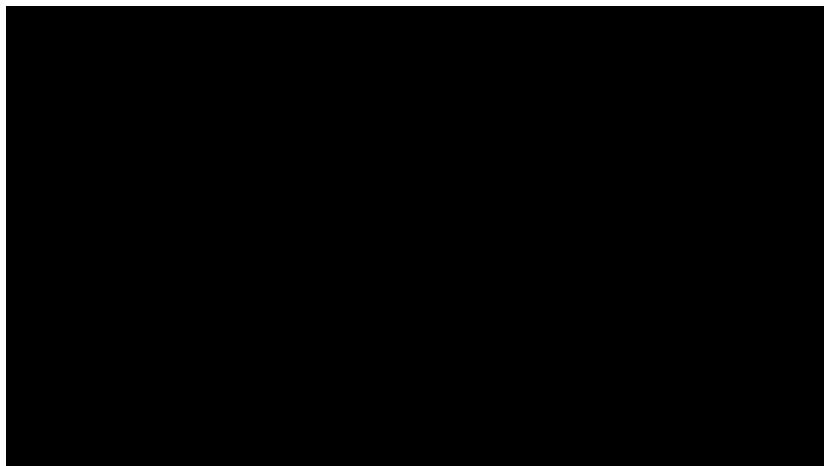
(Srinivasan et al.
2017)

Application: Understand the model



Motion vectors can be extracted
from the compressed video
-> allows very efficient analysis

- Fisher Vector / SVM classifier
- Model of Kantorov & Laptev, (CVPR'14)
- Histogram Of Flow, Motion Boundary Histogram
- HMDB51 dataset



(Srinivasan et al.
2017)

Application: Understand the model



movie review:
++, —

- bidirectional LSTM model (Li'16)
- Stanford Sentiment Treebank dataset

How to handle multiplicative interactions ?

$$z_j = z_g \cdot z_s$$

$$R_g = 0 \quad R_s = R_j$$

gate neuron indirectly affect relevance distribution in forward pass

Negative sentiment

... too slow , too boring , and occasionally annoying .

it 's neither as romantic nor as thrilling as it should be .

neither funny nor suspenseful nor particularly well-drawn .

Model understands negation !

(Arras et al., 2017 & 2018)

Application: Understand the model

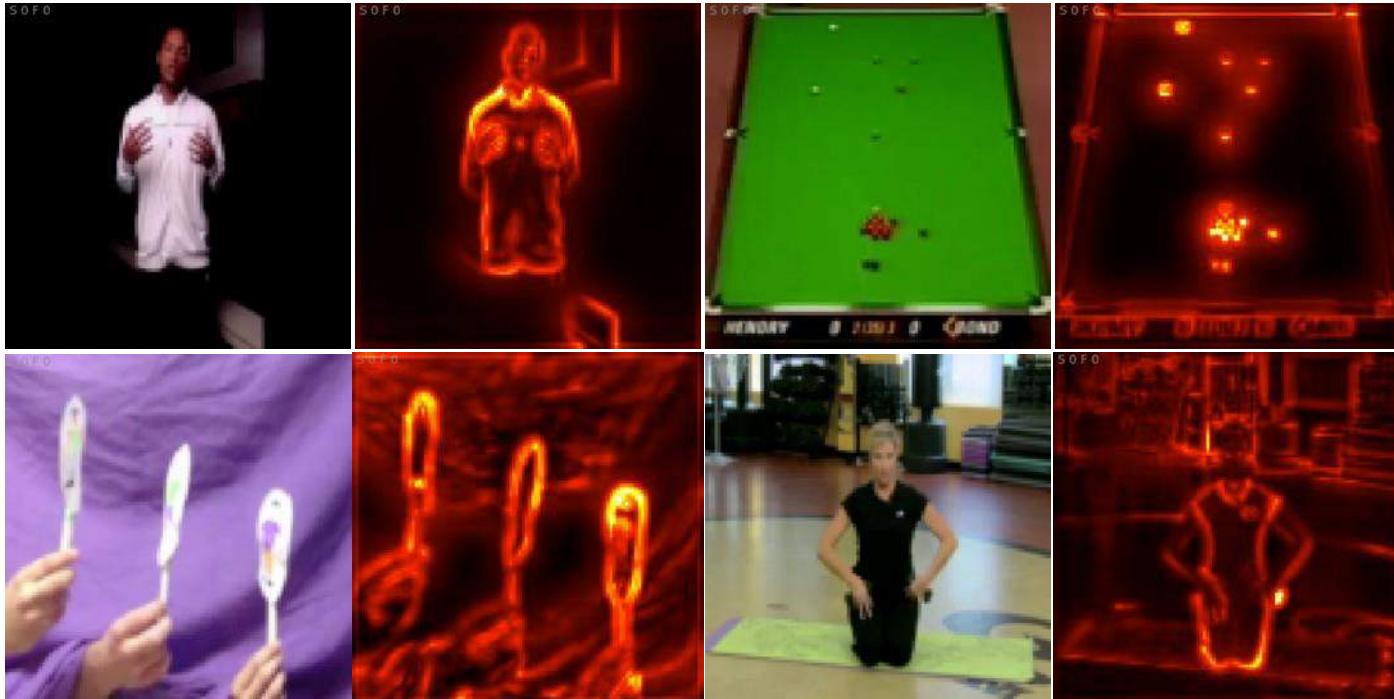
- 3-dimensional CNN (C3D)
- trained on Sports-1M
- explain predictions for 1000 videos from the test set

frame 1 frame 4 frame 7 frame 10 frame 13 frame 16



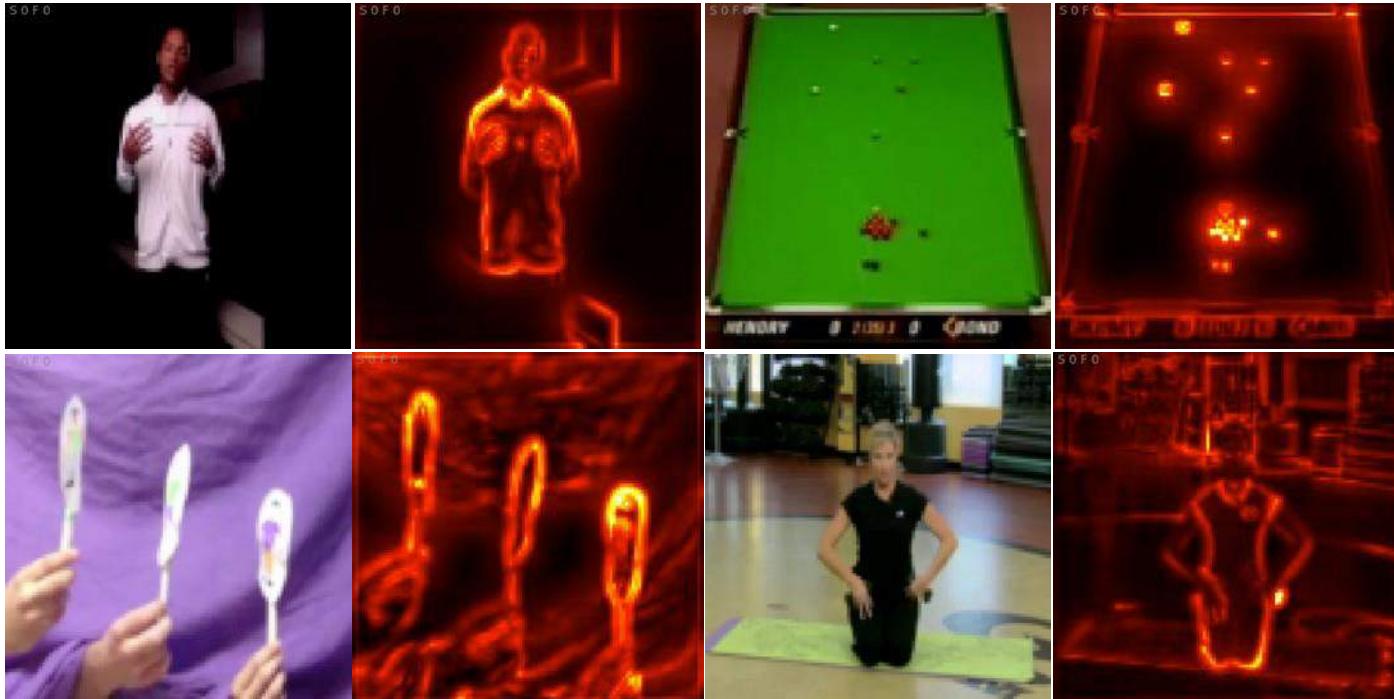
(*Anders et al.,
2018*)

Application: Understand the model



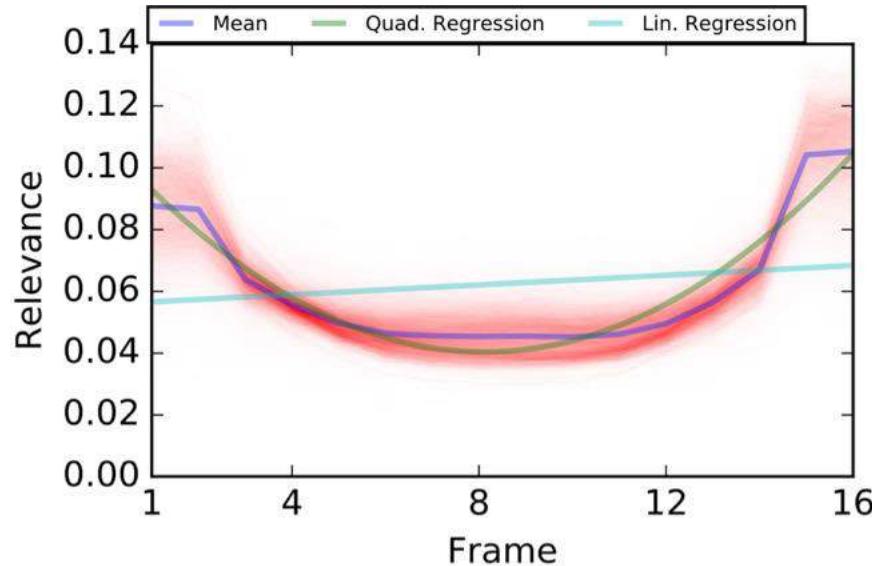
(Anders et al.,
2018)

Application: Understand the model



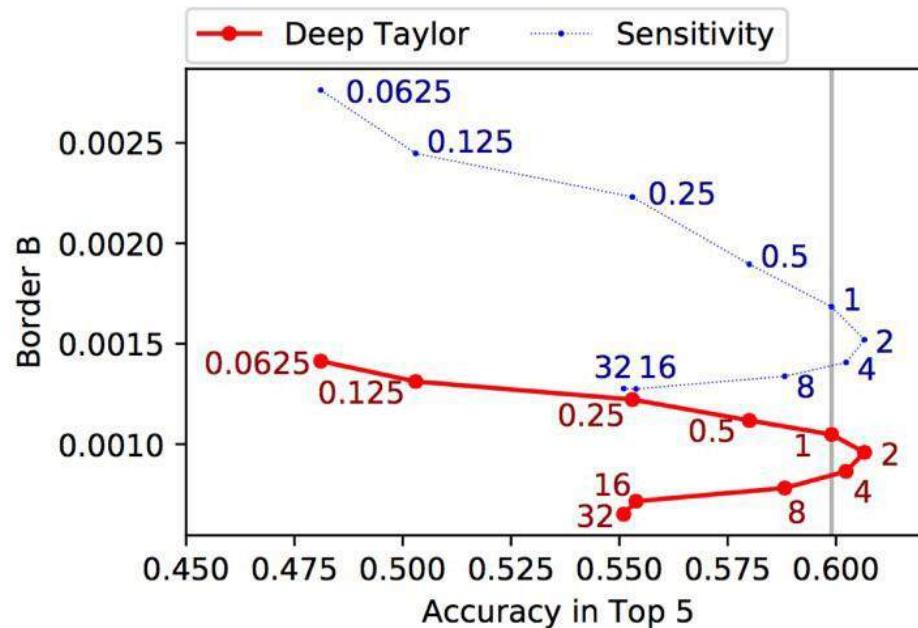
(Anders et al.,
2018)

Application: Understand the model



Observation: Explanations focus on the bordering of the video, as if it wants to watch more of it.

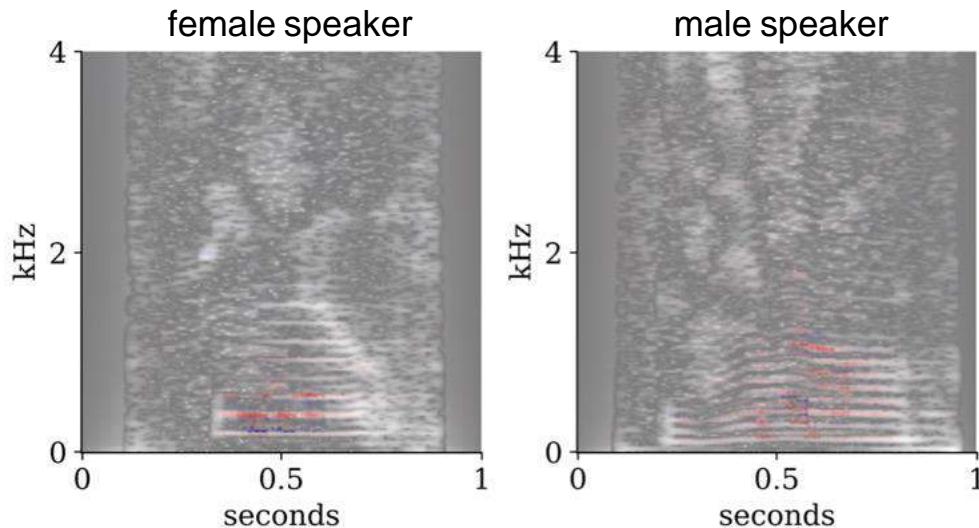
Application: Understand the model



Idea: Play video in fast forward (without retraining) and then the classification accuracy improves.

Application: Understand the model

- AlexNet model
- trained on spectrograms
- spoken digits dataset (AudioMNIST)



model classifies gender based on the fundamental frequency and its immediate harmonics (see also Traunmüller & Eriksson 1995)

(Becker et al.,
2018)

Application: Understand the model

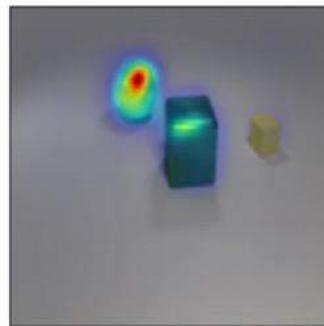
Question

there is a metallic cube ; are
there any large cyan metallic
objects behind it ?



LRP

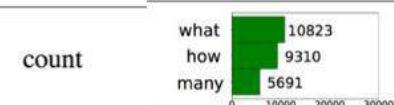
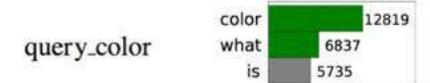
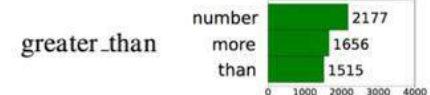
there is a metallic cube ; are
there any large cyan metallic
objects behind it ?



- reimplement model of (Santoro et al., 2017)
- test accuracy of 91,0%
- CLEVR dataset

Question Type

LRP

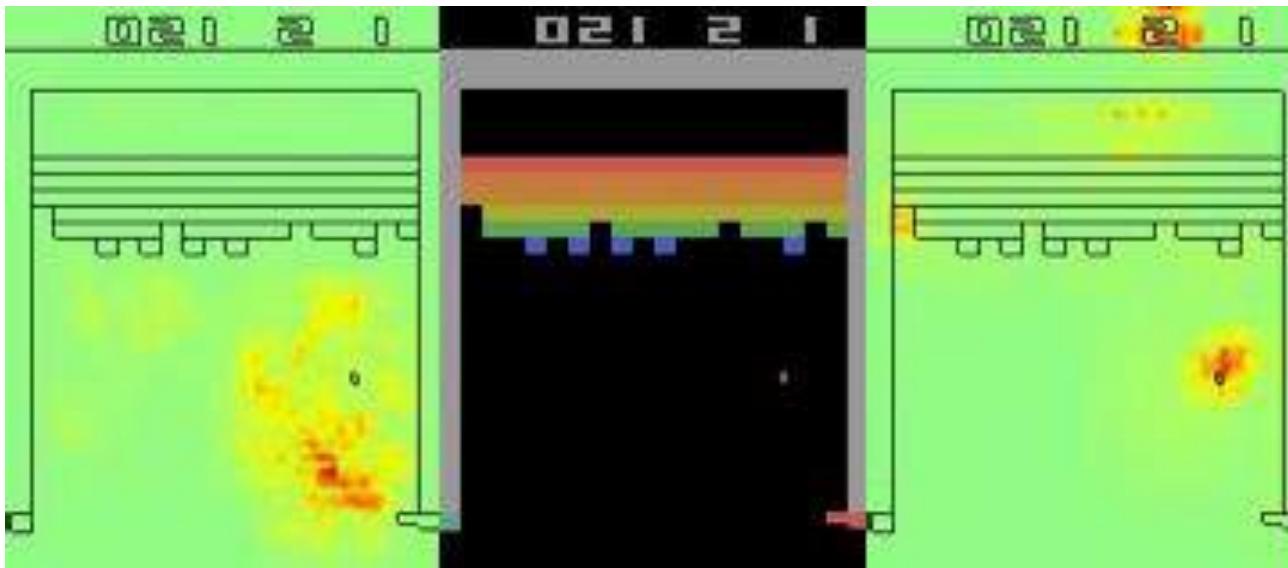


model understands the question and correctly identifies
the object of interest

(Arras et al.,
2018)

Application: Understand the model

Sensitivity Analysis

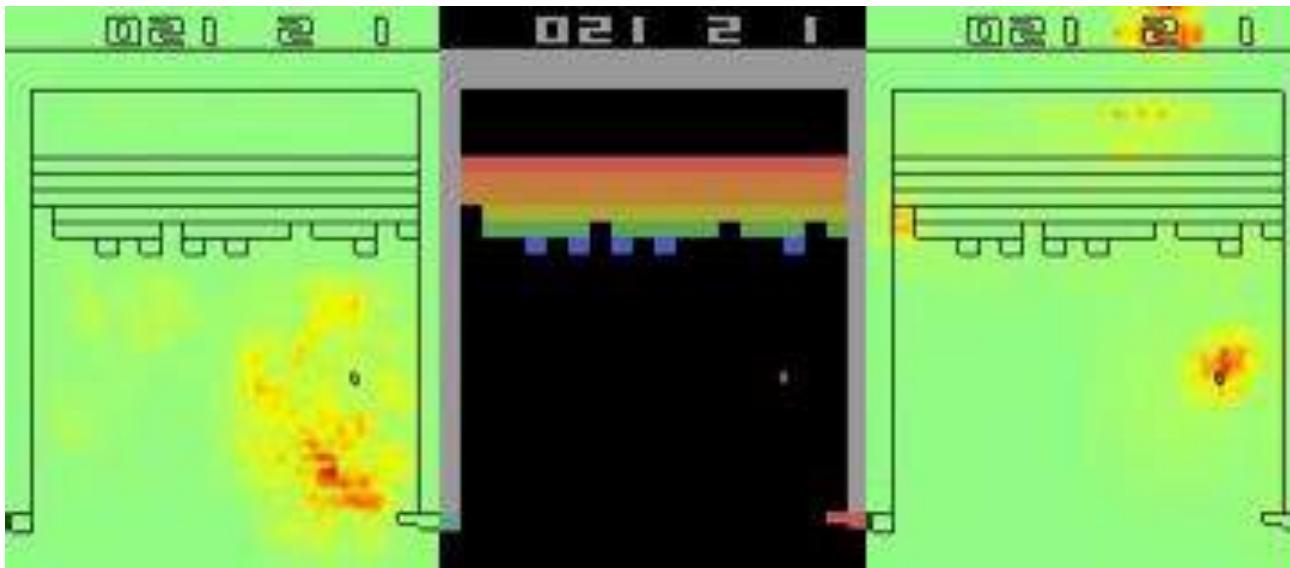


does not focus on where the ball is, but on where the ball could be in the next frame

LRP shows that that model tracks the ball
(Lapuschkin et al., in prep.)

Application: Understand the model

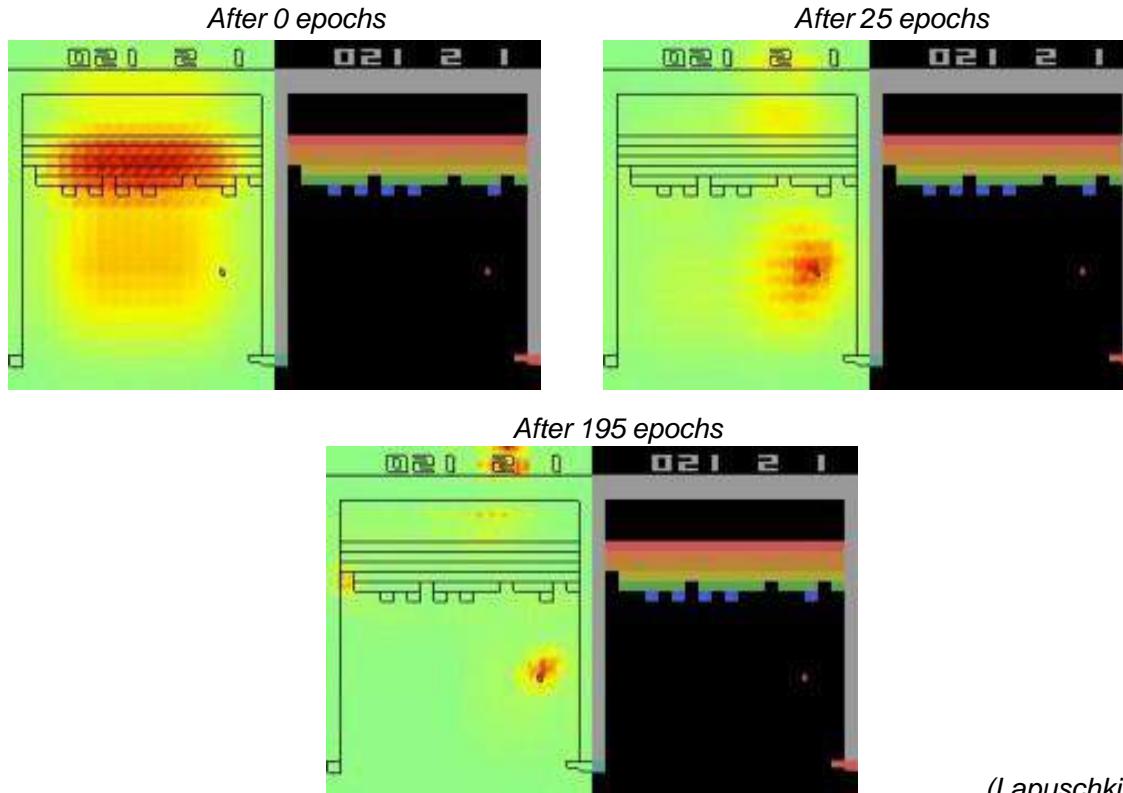
Sensitivity Analysis



does not focus on where the ball is, but on where the ball could be in the next frame

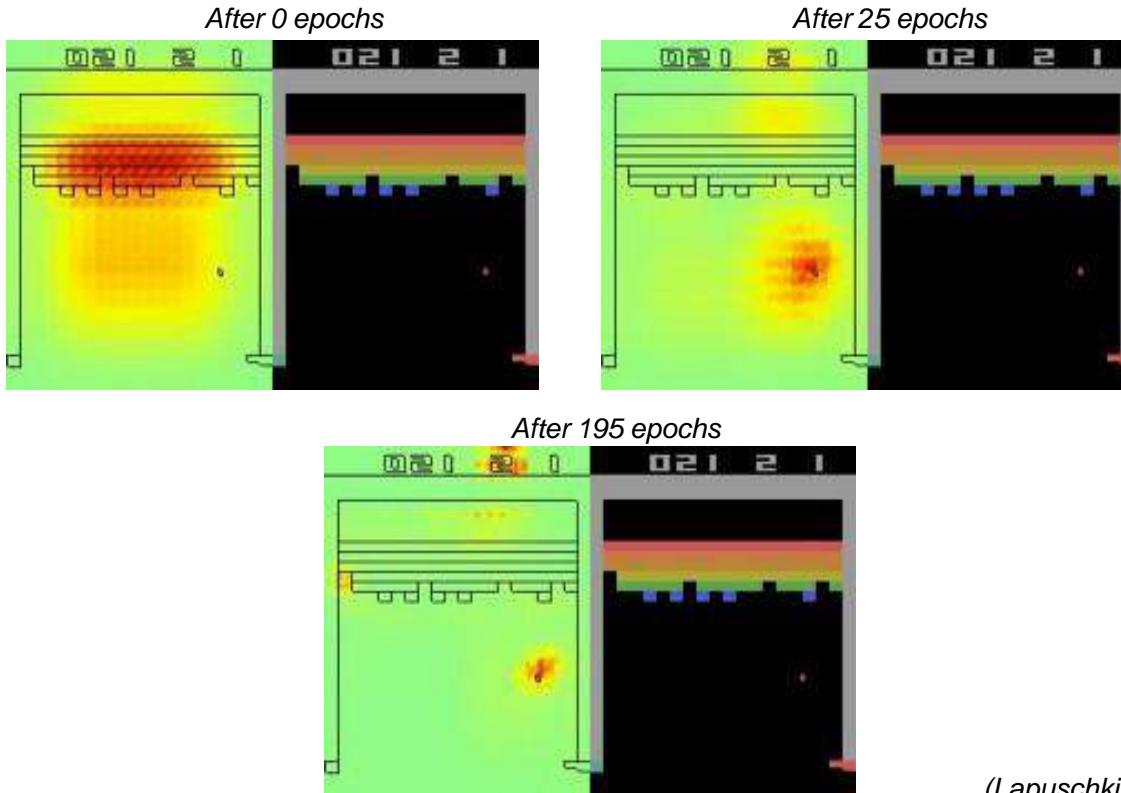
LRP shows that that model tracks the ball
(Lapuschkin et al., in prep.)

Application: Understand the model



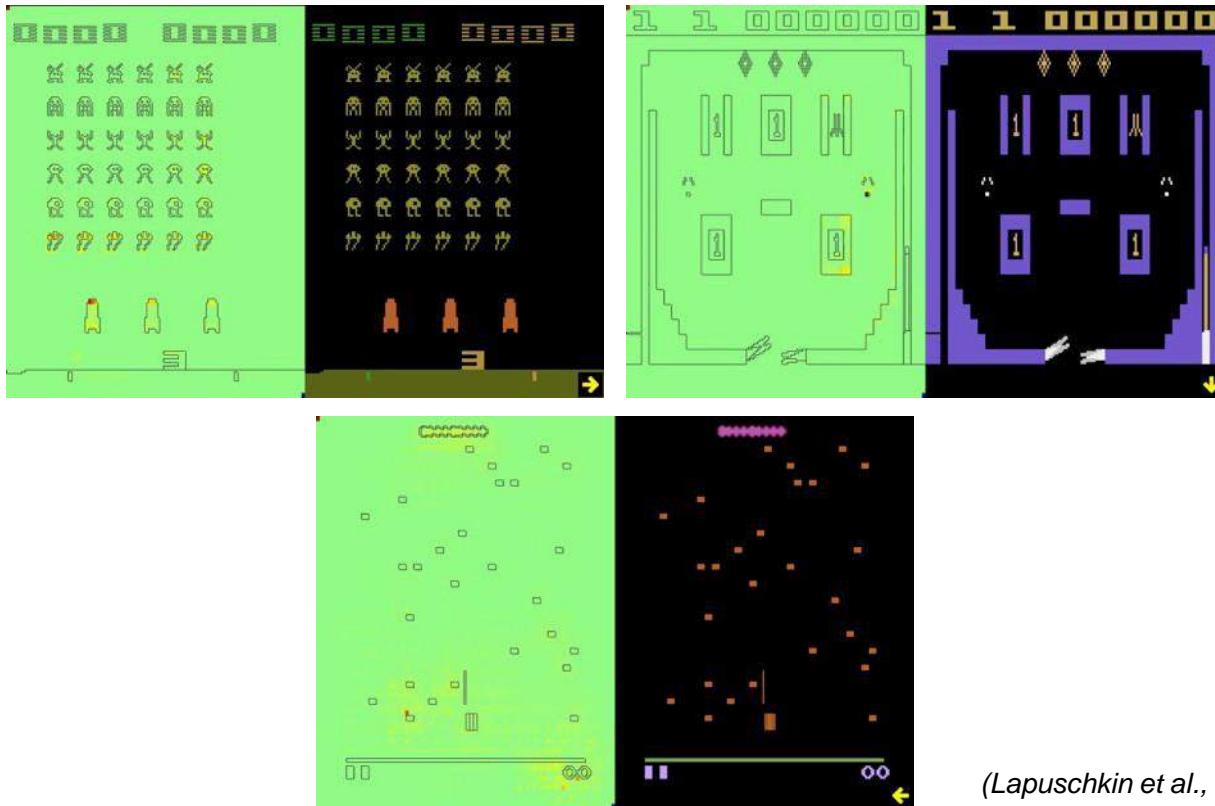
(Lapuschkin et al., in prep.)

Application: Understand the model



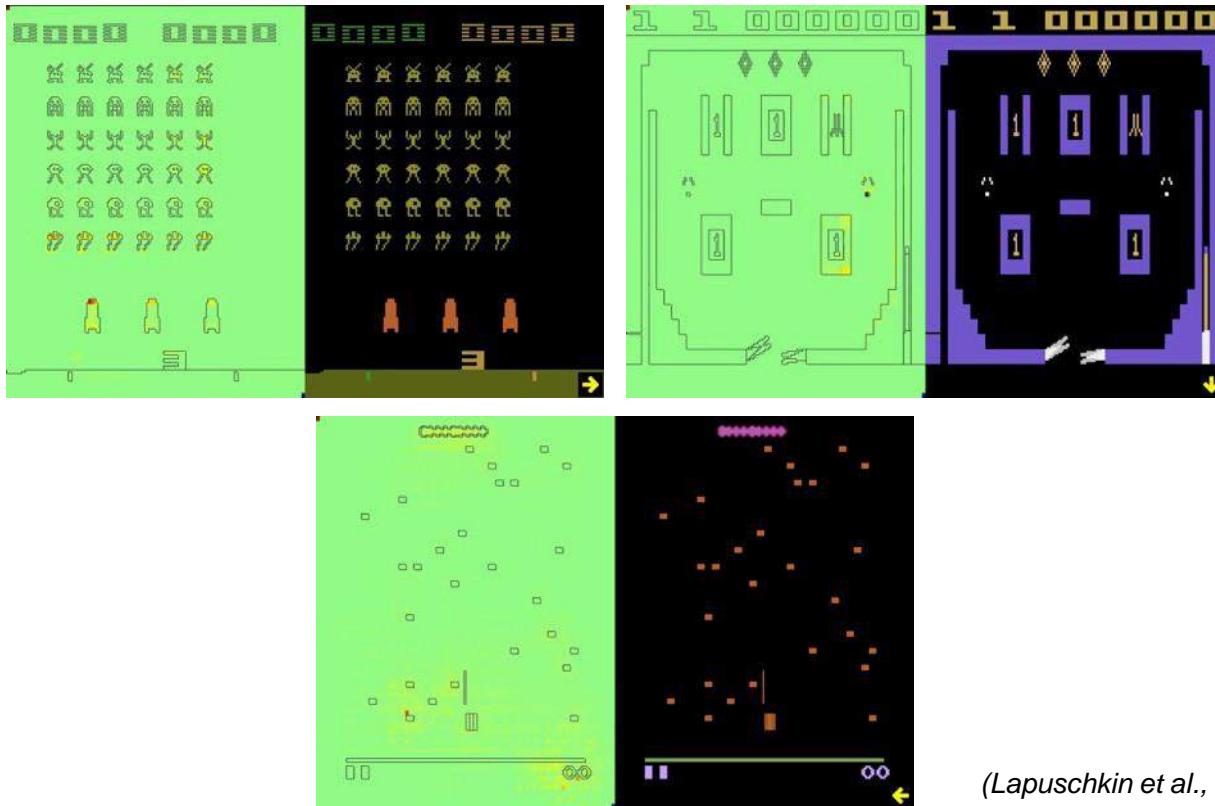
(Lapuschkin et al., in prep.)

Application: Understand the model



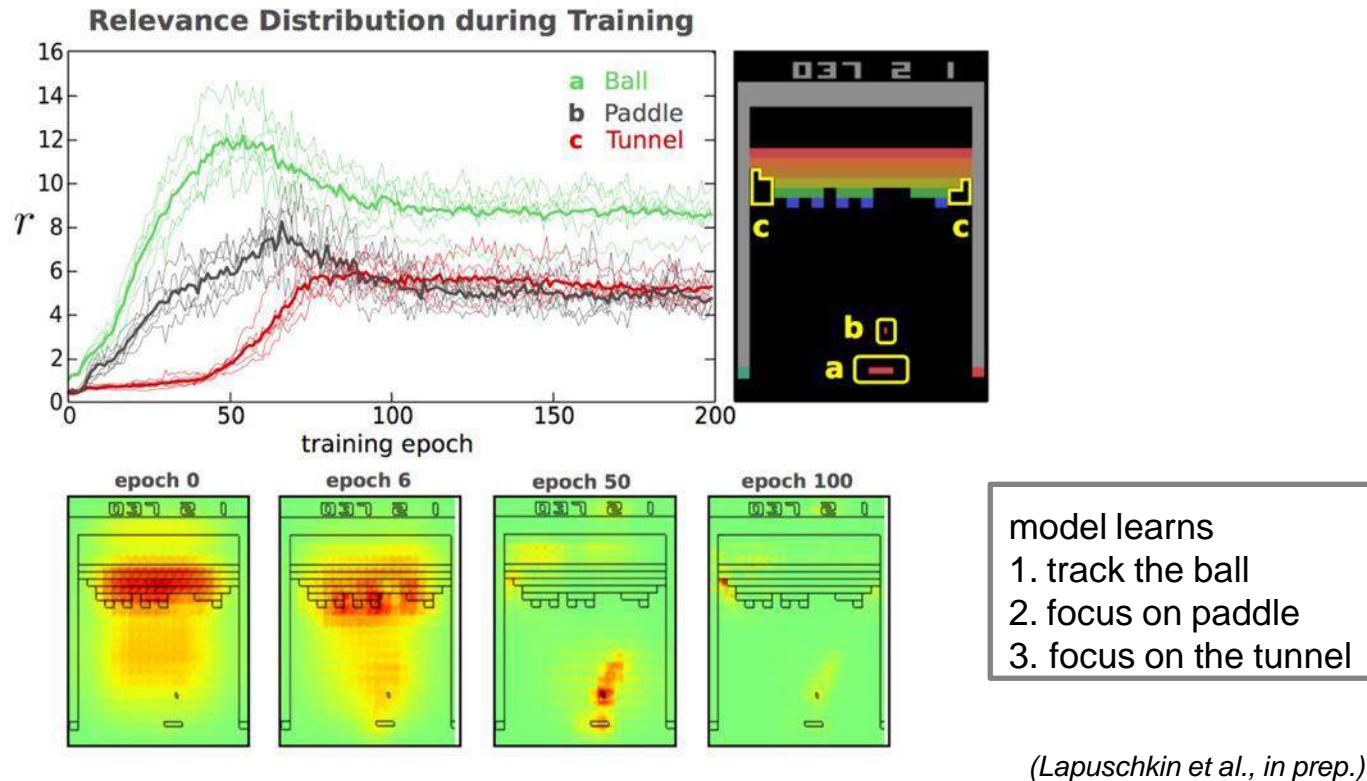
(Lapuschkin et al., in prep.)

Application: Understand the model



(Lapuschkin et al., in prep.)

Application: Understand the model

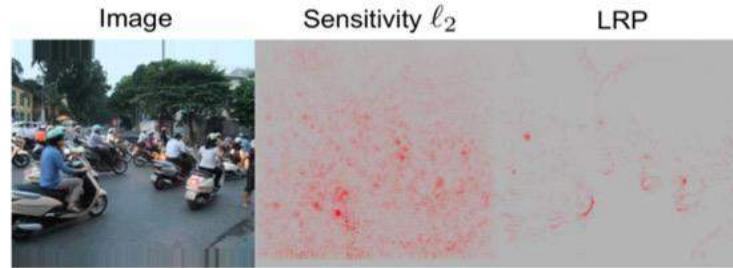


Tutorial on Interpretable Machine Learning

Wrap-up

Take Home Messages

Sensitivity analysis is not the question that you would like to ask!



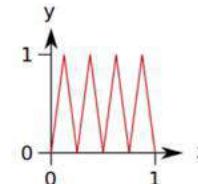
Take Home Messages

What works for simple models doesn't work for deep models.

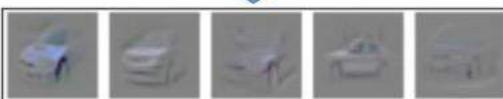


gradient-based methods

vulnerable to shattered gradients



Our LRP method is robust to this.



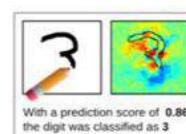
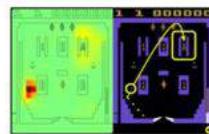
Take Home Messages

LRP works 4 all: deep models, LSTMs, kernel methods ...



LRP Explanation Framework

e people are more prone to g
The mental part is usually
y is up or down, ie: the Shu
oointed towards Earth, so the
astronauts. About 50% of t
s, and NASA has done numerou



(software, tutorials, demos,
insights, applications)



Take Home Messages

LRP \neq Gradient \times Input

... except for special cases. LRP was developed among others because gradient-based methods aren't satisfying.

High flexibility: Different LRP variants, free parameters

Good news: No need to reimplement LRP, check our software at www.heatmapping.org.

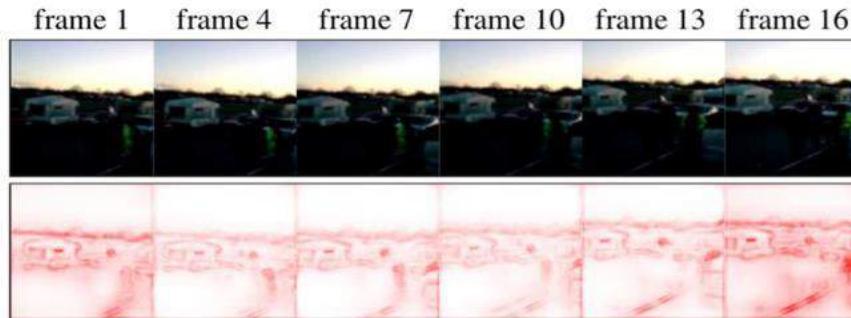
Take Home Messages

Explanations can be evaluated:
Pixel flipping (model agnostic)
And beyond LRP and DTD

[Samek et al. IEEE TNNLS 2017]

Take Home Messages

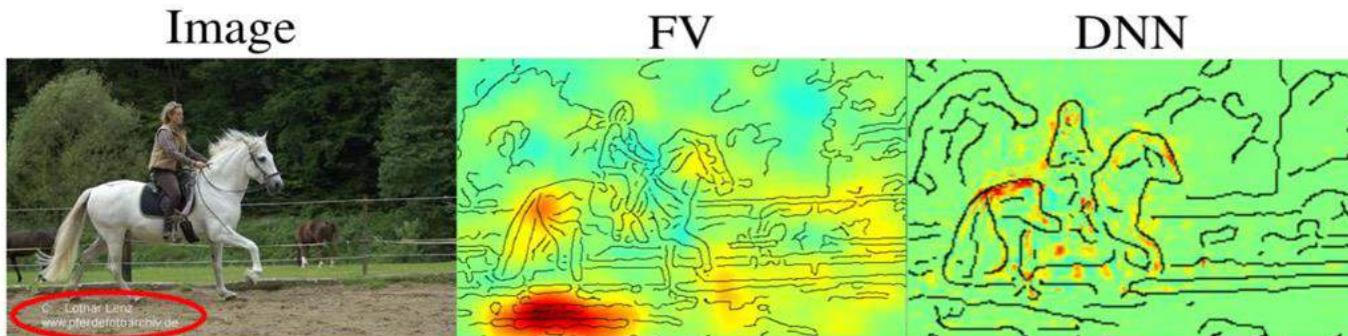
Explanation helps to improve models



Explaining ML, Now What?

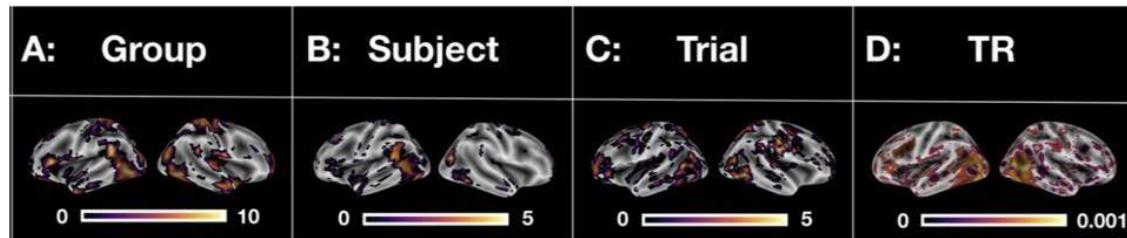
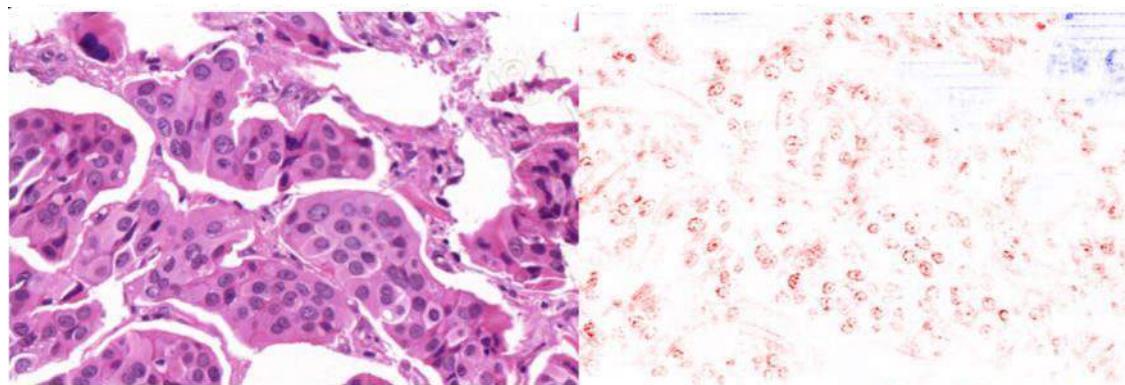
Take Home Messages

Explanation helps to find flaws
in models



Take Home Messages

Getting **new** Insights in the Sciences

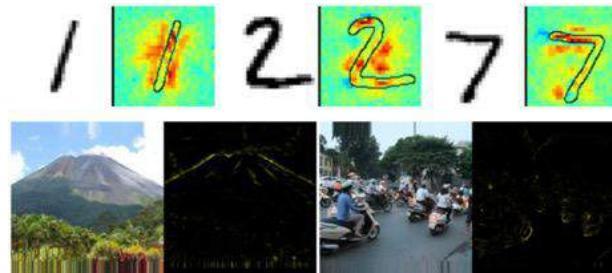


More information

Visit:

<http://www.heatmapping.org>

- ▶ Tutorials
- ▶ Software
- ▶ Online Demos



Tutorial Paper

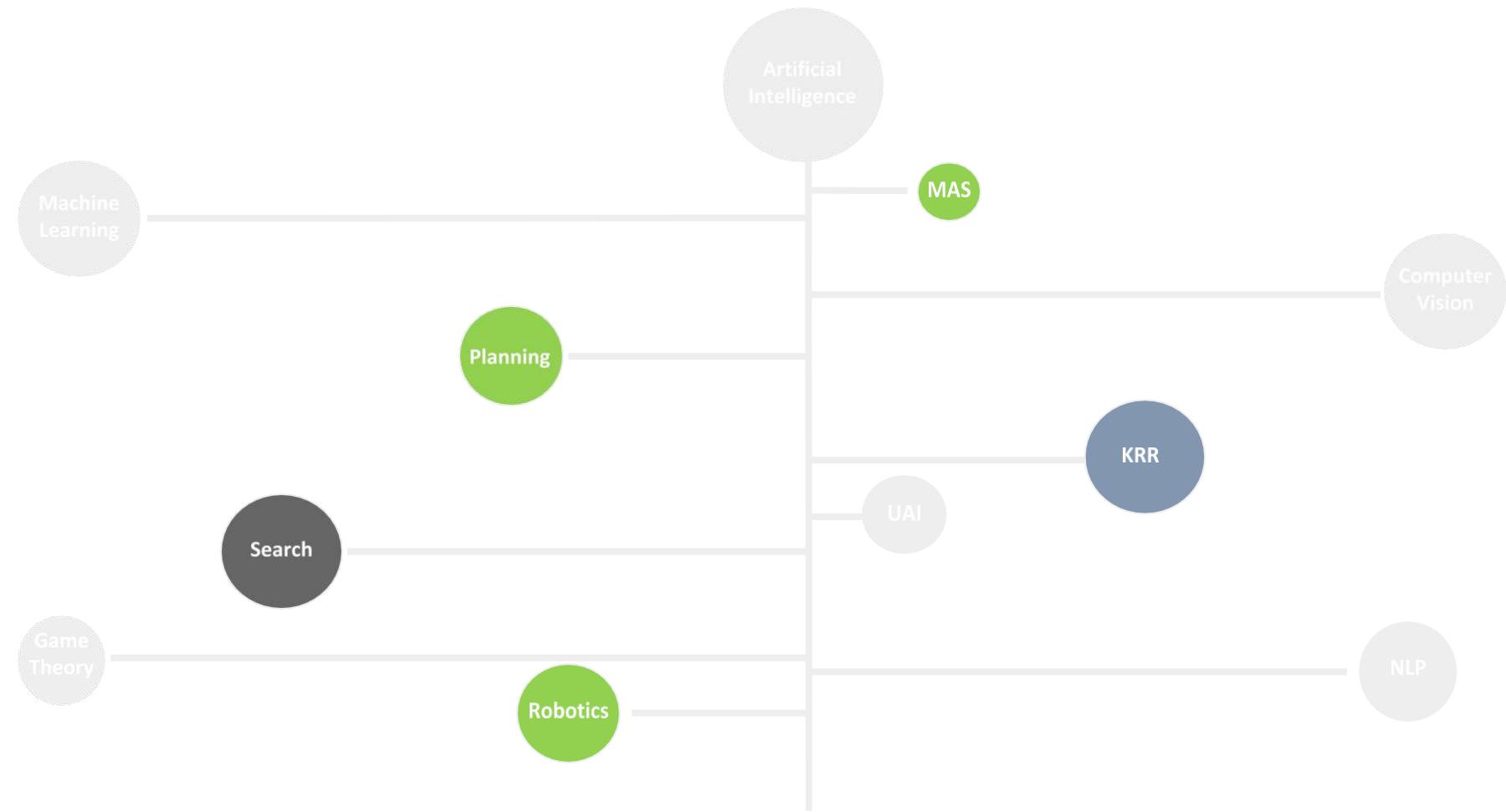
Montavon et al., "Methods for interpreting and understanding deep neural networks",
Digital Signal Processing, 73:1-5, 2018

Keras Explanation Toolbox

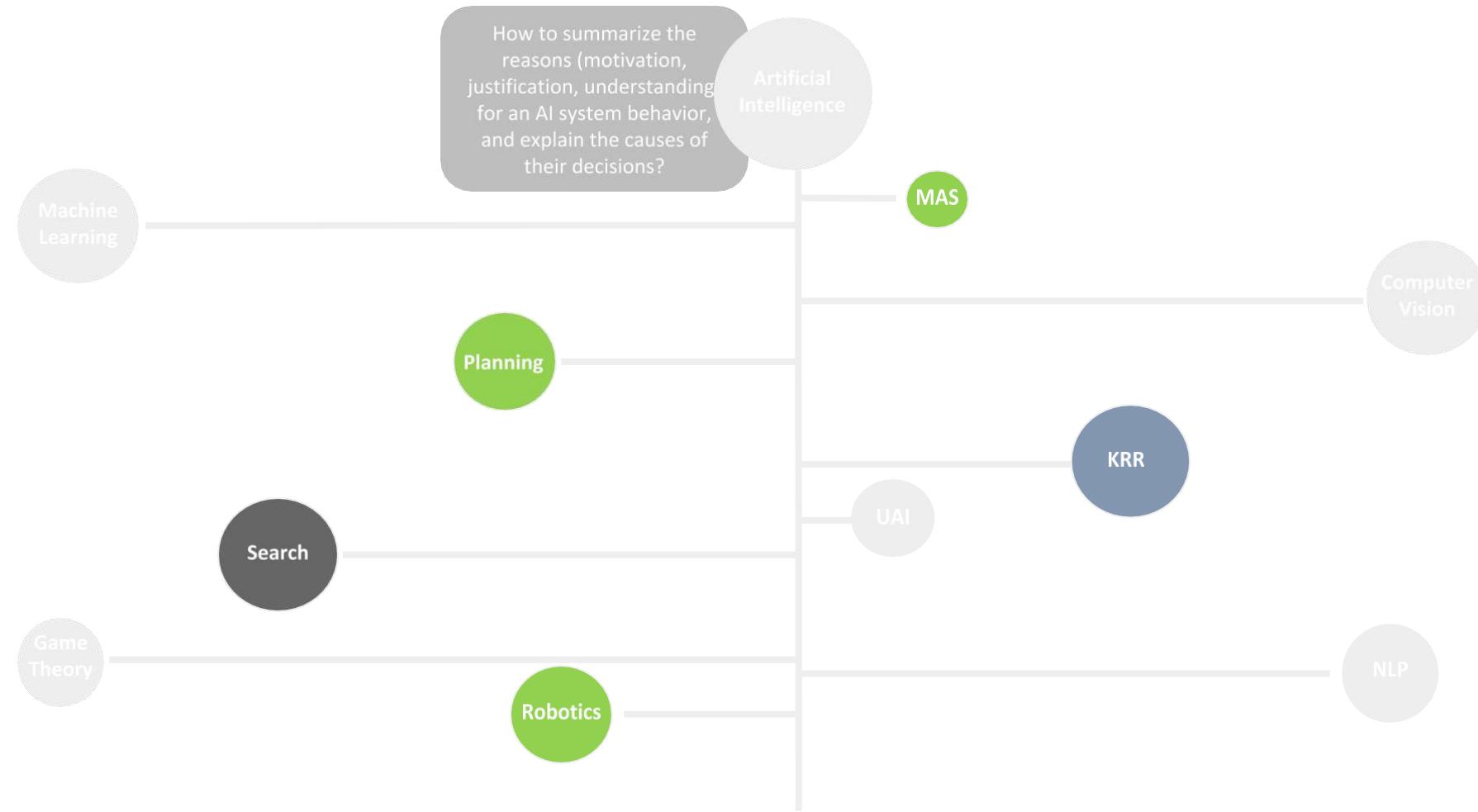
<https://github.com/albermax/innvestigate>

Explanation in AI (not only Machine Learning!)

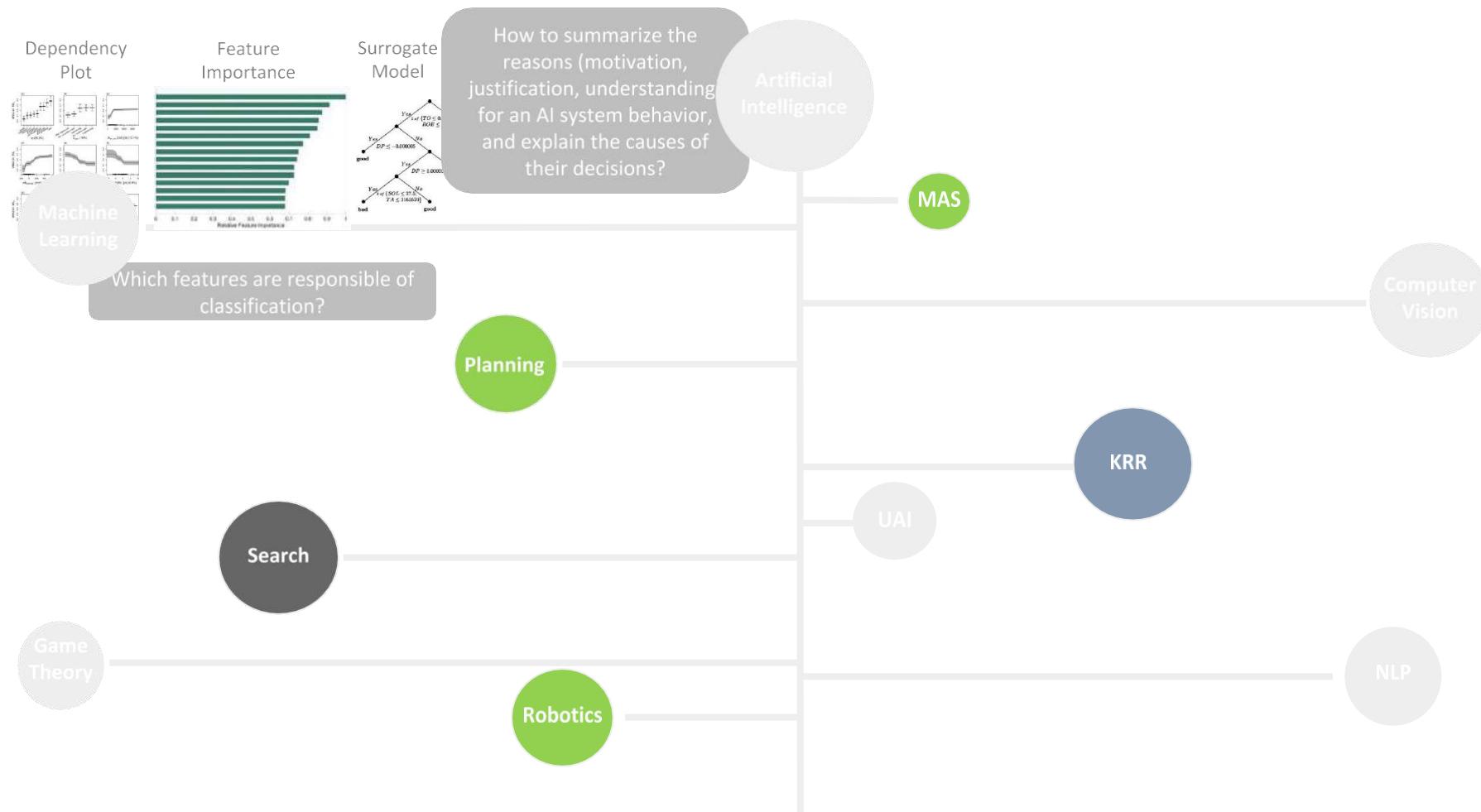
XAI: One Objective, Many ‘AI’s, Many Definitions, Many Approaches



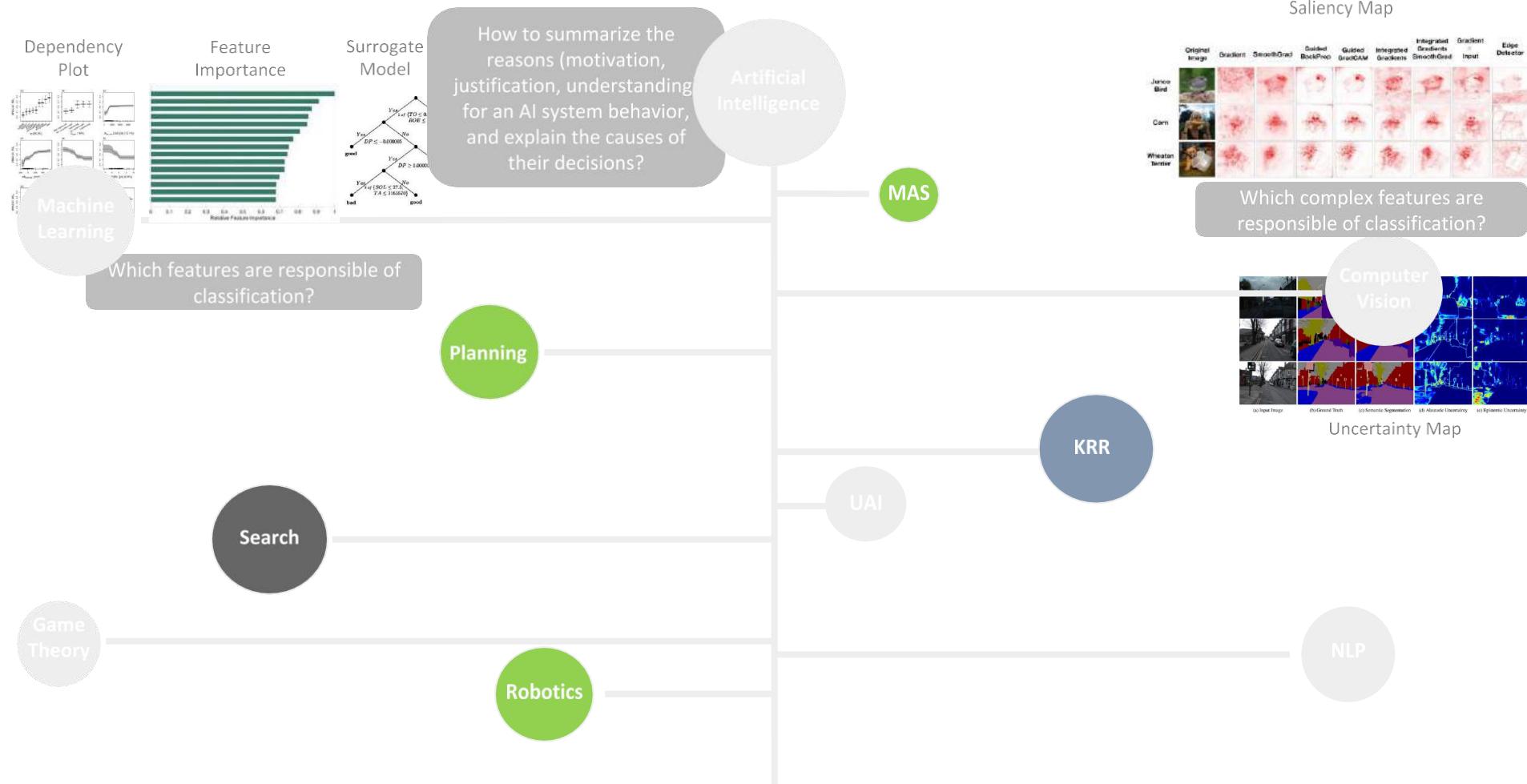
XAI: One Objective, Many ‘AI’s, Many Definitions, Many Approaches



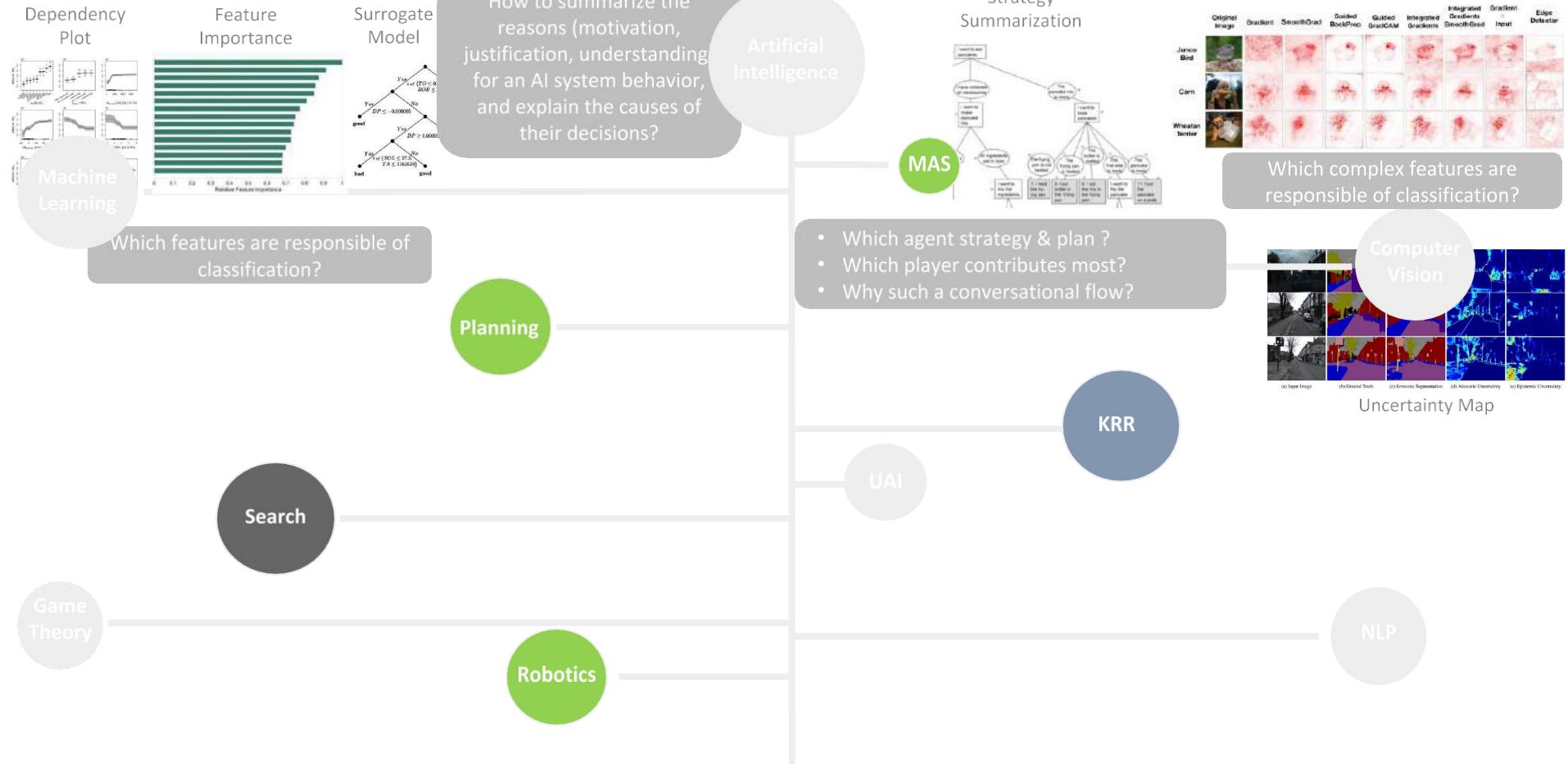
XAI: One Objective, Many ‘AI’s, Many Definitions, Many Approaches



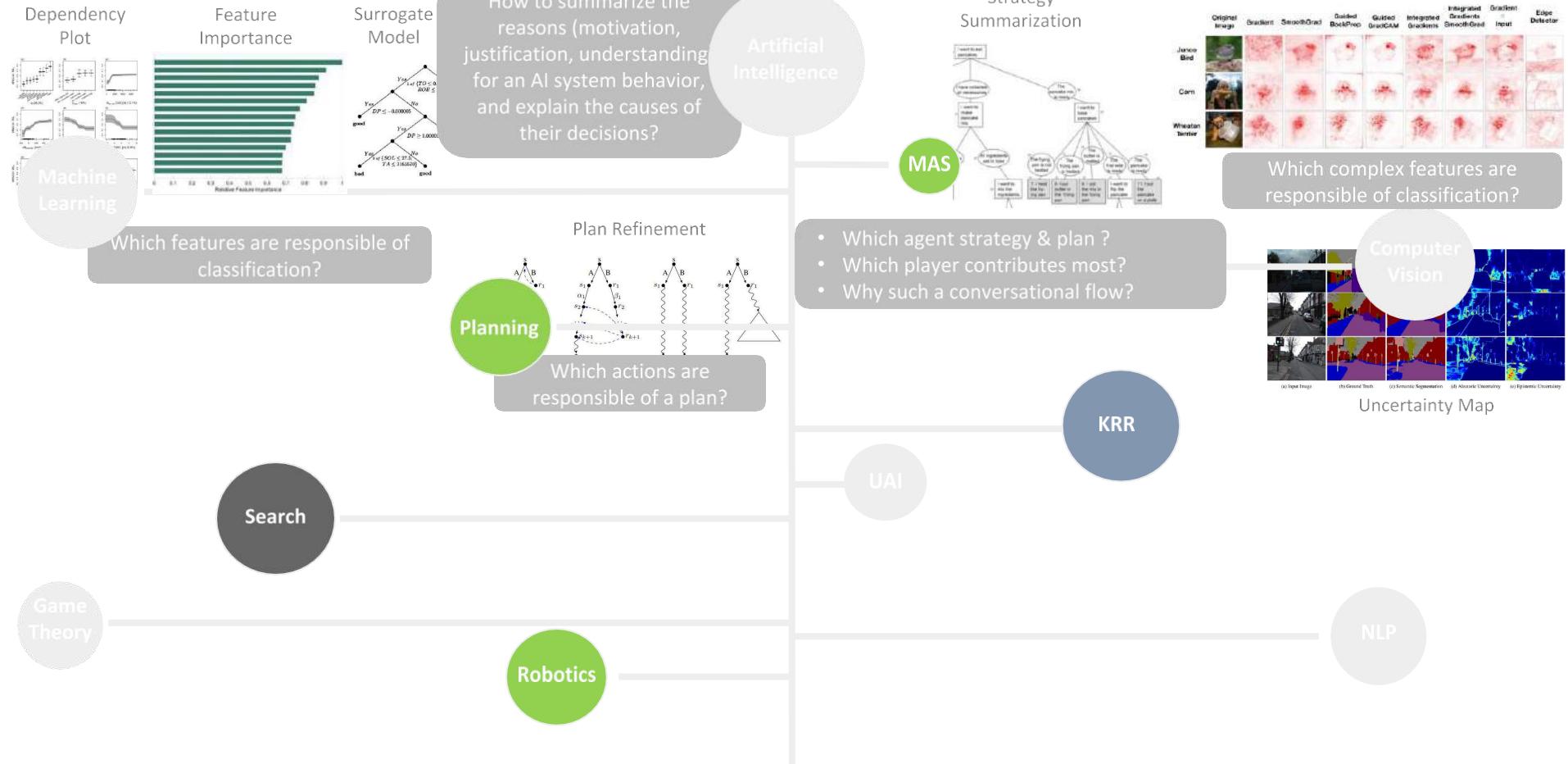
XAI: One Objective, Many ‘AI’s, Many Definitions, Many Approaches



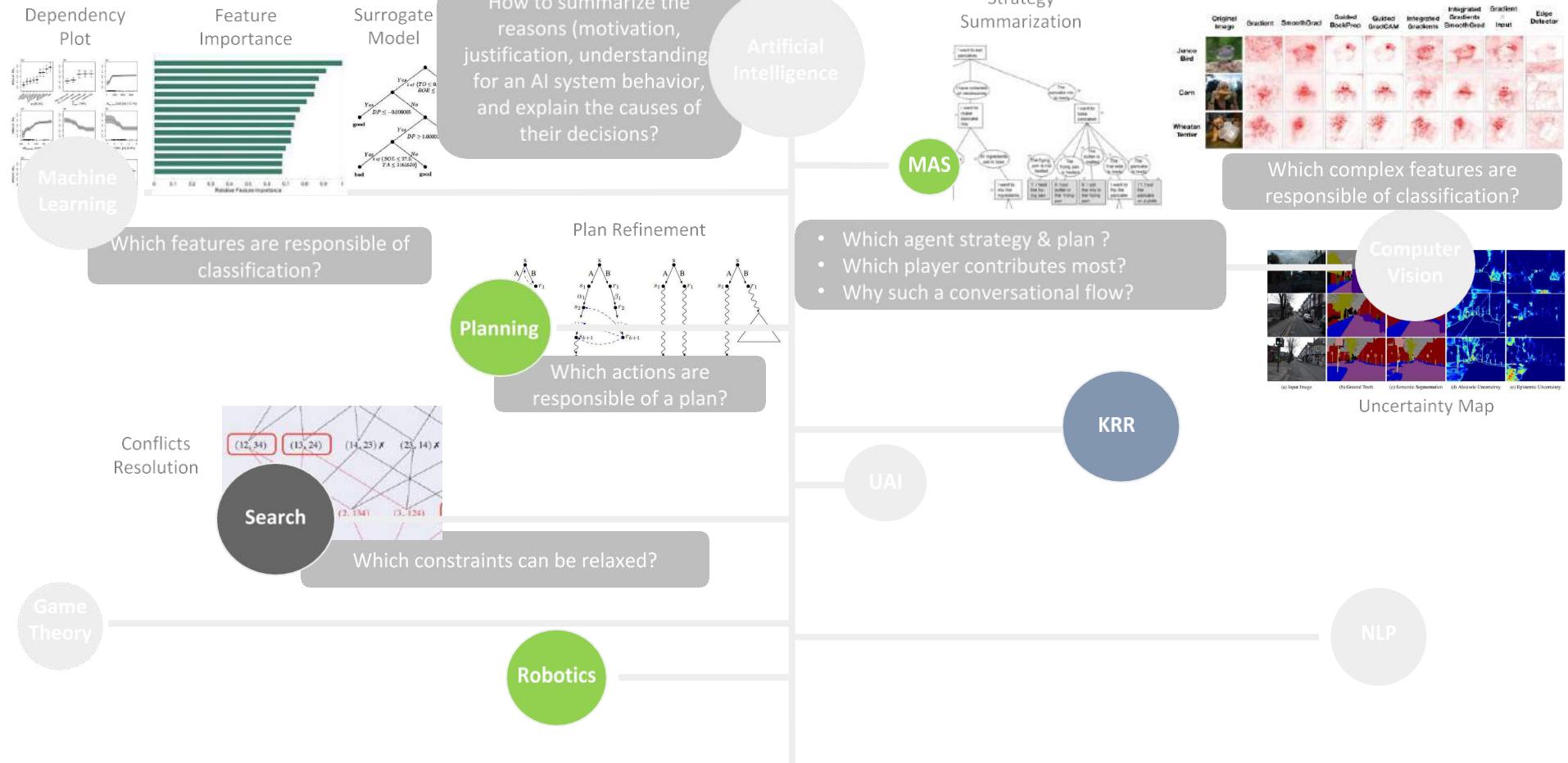
XAI: One Objective, Many ‘AI’s, Many Definitions, Many Approaches



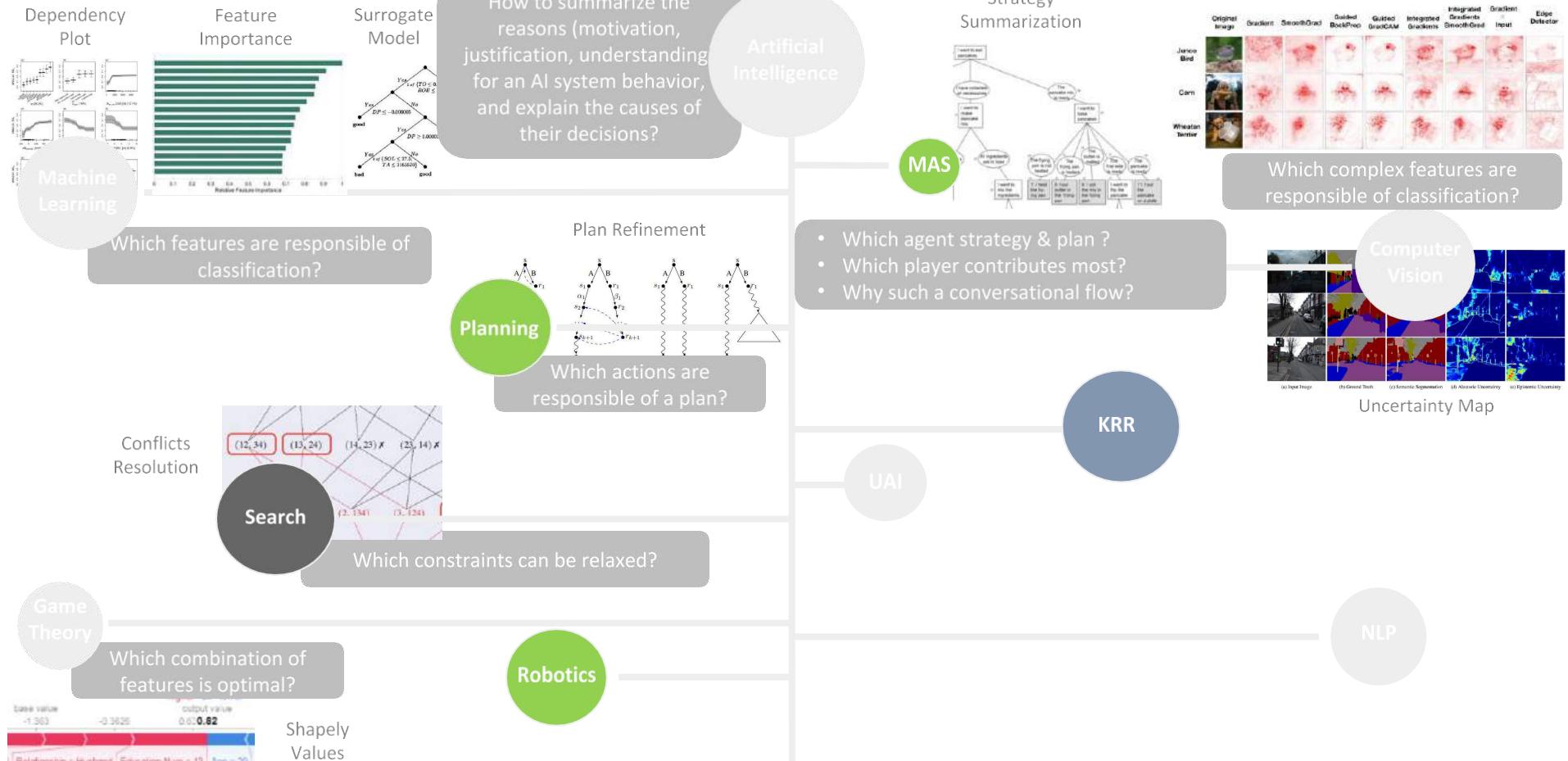
XAI: One Objective, Many ‘AI’s, Many Definitions, Many Approaches



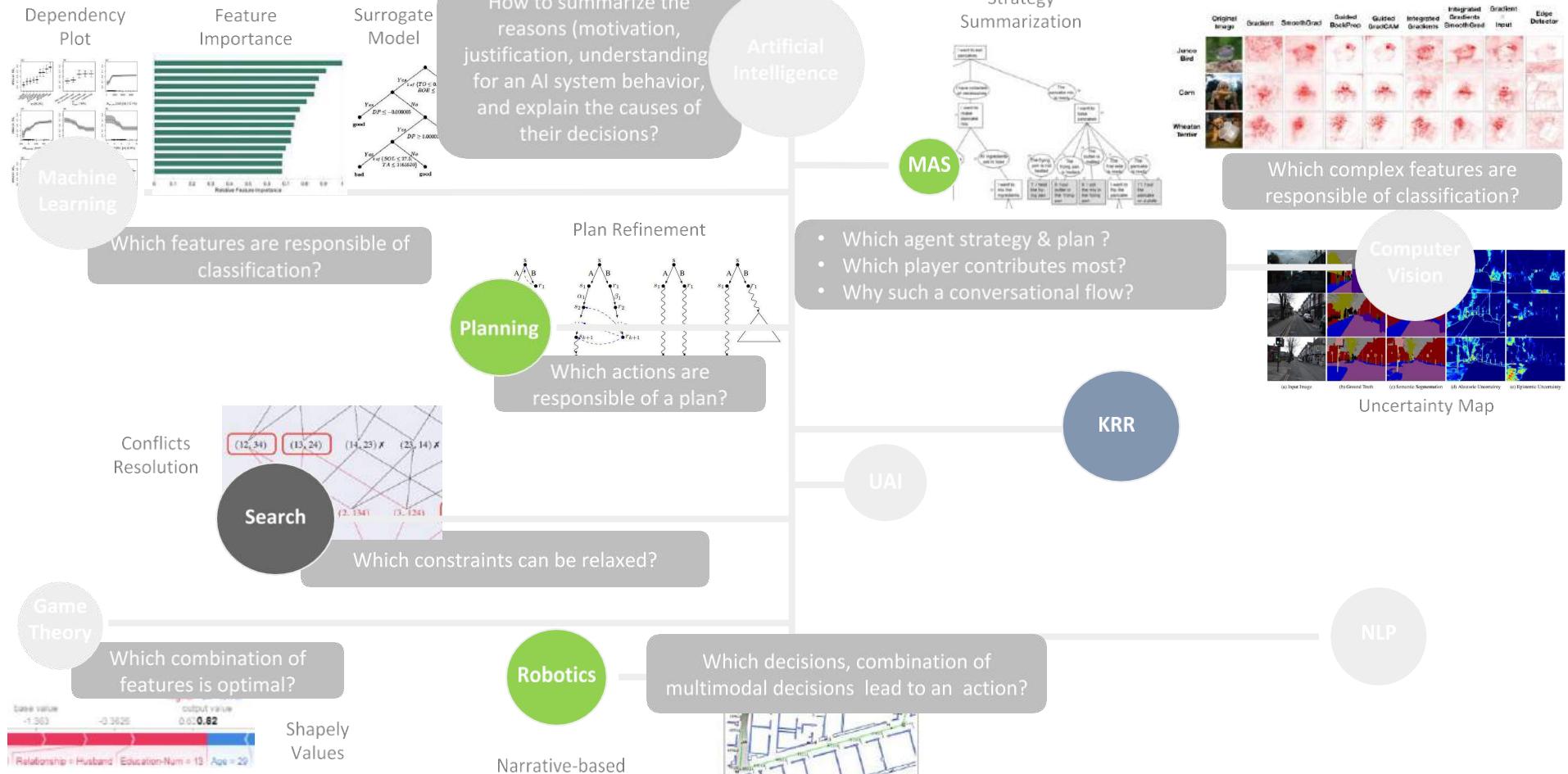
XAI: One Objective, Many ‘AI’s, Many Definitions, Many Approaches



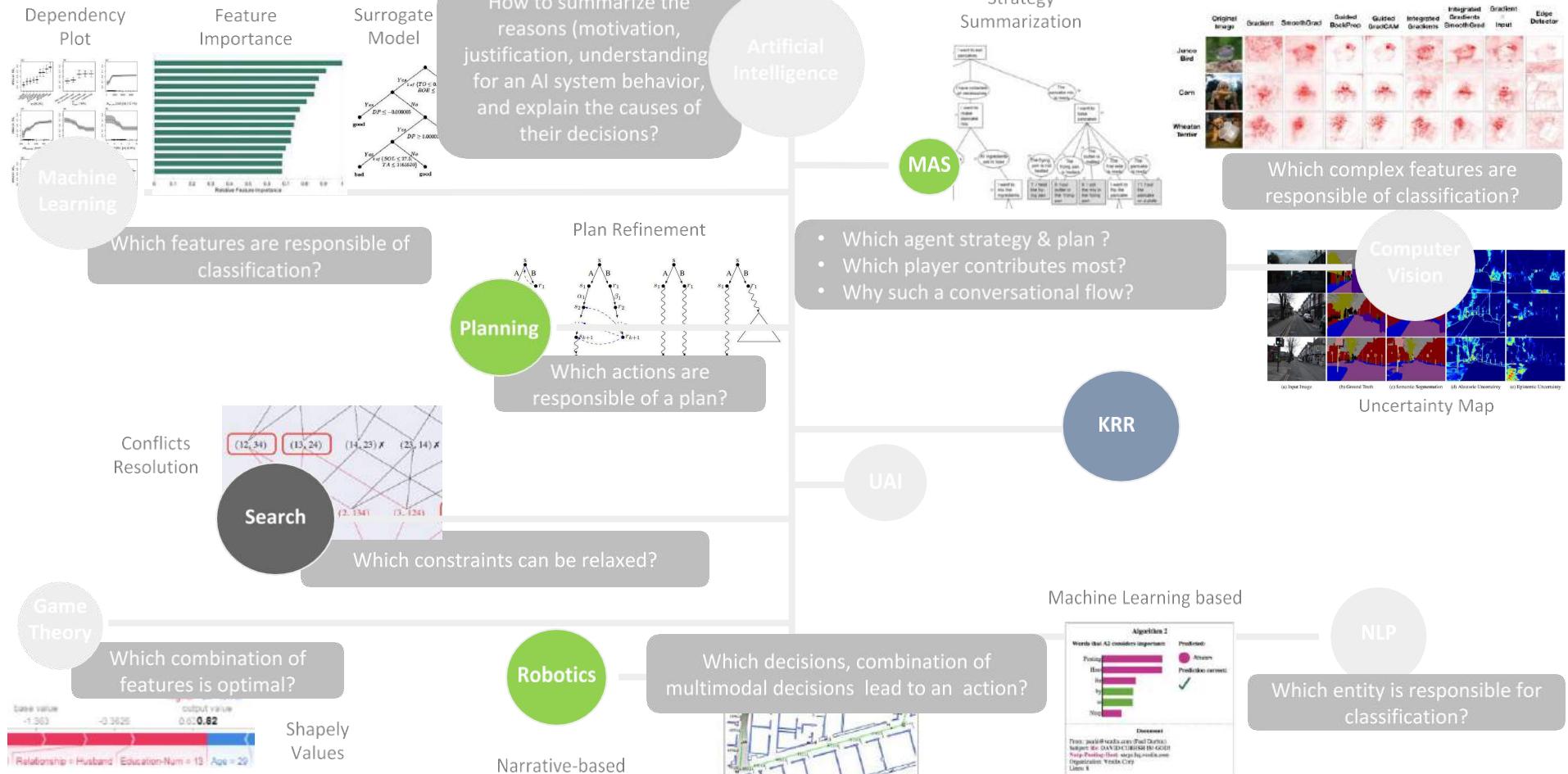
XAI: One Objective, Many ‘AI’s, Many Definitions, Many Approaches



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



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



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

Machine Learning

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?)

Which features are responsible of classification?

Planning

Plan Refinement

Which actions are responsible of a plan?

Search

Conflicts Resolution

Which constraints can be relaxed?

Robotics

Narrative-based

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

Game Theory

Shapely Values

Which combination of features is optimal?

Artificial Intelligence

Strategy Summarization

MAS

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

Computer Vision

Saliency Map

Original Image, Gradient, SmoothGrad, Guided BackProp, Guided GradCAM, Integrated Gradients, Integrated Gradients SmoothGrad, Gradient Input, Edge Detector, Junco Bird, Corn, Wheat Tassle

Which complex features are responsible of classification?

KRR

UAI

THING
AndR
AddL
All
Allt
Eq

$\vdash \text{at-least } p \equiv (\text{at-least } m)$

$\vdash C \equiv (\text{and } C)$

$\vdash (\text{at-least } p) \equiv \text{THP}$

$\vdash (\text{all } p \text{ THING}) \equiv r$

Diagnosis

Abduction

Uncertainty Map

(a) Input Image, (b) Ground Truth, (c) Semantic Segmentation, (d) Atomsic Geometry, (e) Epistemic Uncertainty

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

Machine Learning based

Alibi

Words that AI considers important: Playing, Home, Age, Education-Num, Relationship = Husband, Education-Num = 13, Age = 29.

Predicted: Ames, Prediction correct: ✓

Decides:

From: <https://www.kaggle.com/c/house-prices-advanced-regression-techniques>
Model: Random Forest Regressor
Step: Preprocess data step by step
Organization: Yoda Corp
Date: 8

Which entity is responsible for classification?

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

Machine Learning

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?

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

Narrative-based

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

Artificial Intelligence

Strategy Summarization

MAS

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

Computer Vision

Saliency Map

Original Image Gradient SmoothGrad Guided BackProp Guided GradCAM Integrated Gradients Integrated SmoothGrad Gradient Input Edge Detector

Junco Bird Corn Wheat Tassels

Which complex features are responsible of classification?

Diagnosis

Abduction

Uncertainty Map

KRR

UAI

Uncertainty as an alternative to explanation

Machine Learning based

Alibi

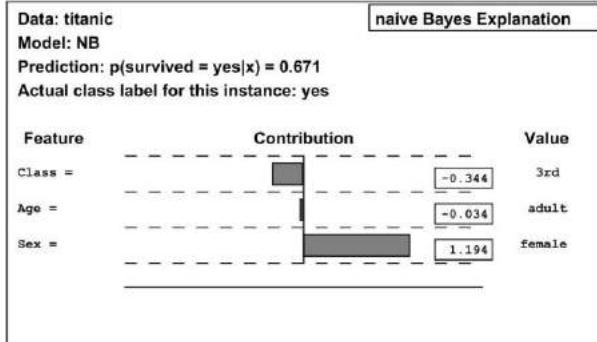
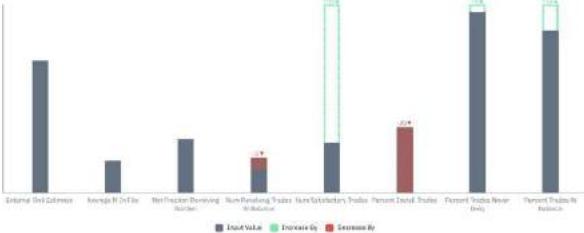
Decoys

Which entity is responsible for classification?

Overview of Explanation in Machine Learning (1)

Interpretable Models:

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



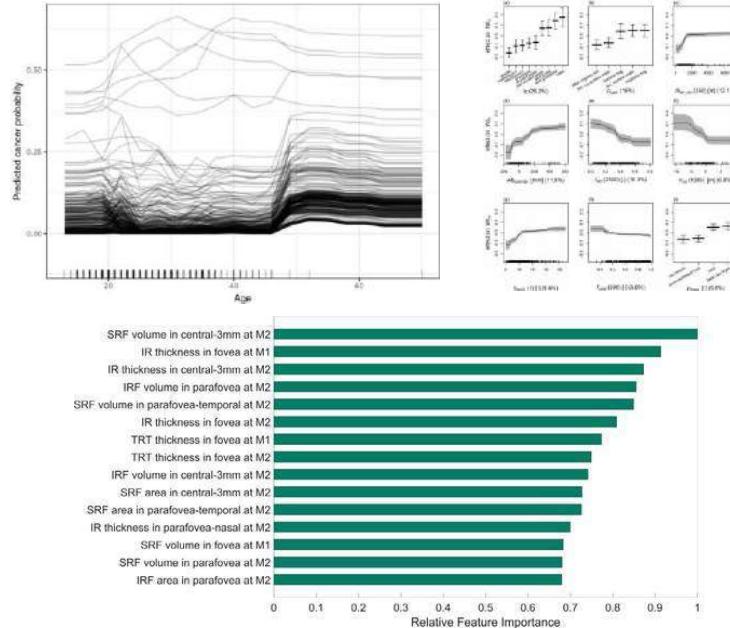
Naive Bayes model

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

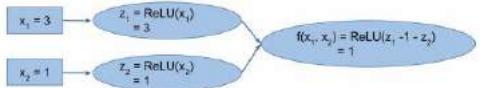
Rory Mc Grath, Luca Costabello,
Chan Le Van, Paul Sweeney,
Farbod Kamiab, Zhao Shen,
Freddy Lécué: Interpretable Credit
Application Predictions With
Counterfactual Explanations.
CoRR abs/1811.05245 (2018)



Feature Importance
Partial Dependence Plot
Individual Conditional Expectation
Sensitivity Analysis

Overview of Explanation in Machine Learning (2)

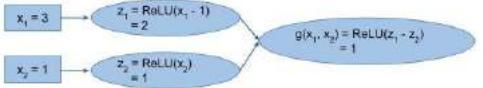
• Artificial Neural Network



Network $f(x_1, x_2)$

Attributions at $x_1 = 3, x_2 = 1$

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



Network $g(x_1, x_2)$

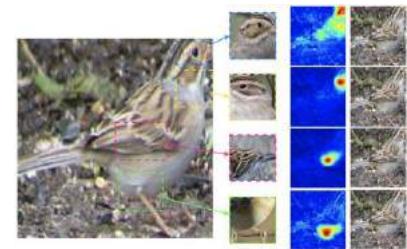
Attributions at $x_1 = 3, x_2 = 1$

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

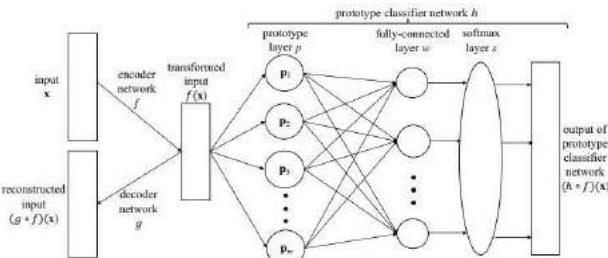
Attribution for Deep Network (Integrated gradient-based)

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

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

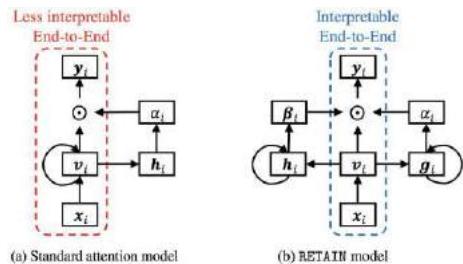


Chaofan Chen, Oscar Li, Alina Barnett, Jonathan Su, Cynthia Rudin: This looks like that: deep learning for interpretable image recognition. CoRR abs/1806.10574 (2018)



Auto-encoder / Prototype

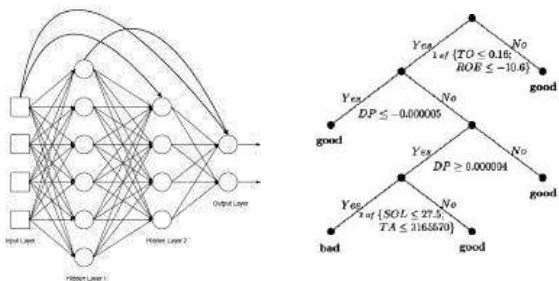
Oscar Li, Hao Liu, Chaofan Chen, Cynthia Rudin: Deep Learning for Case-Based Reasoning Through Prototypes: A Neural Network That Explains Its Predictions. AAAI 2018: 3530-3537



Attention Mechanism

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

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

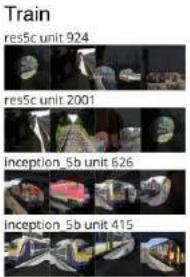


Surrogate Model

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

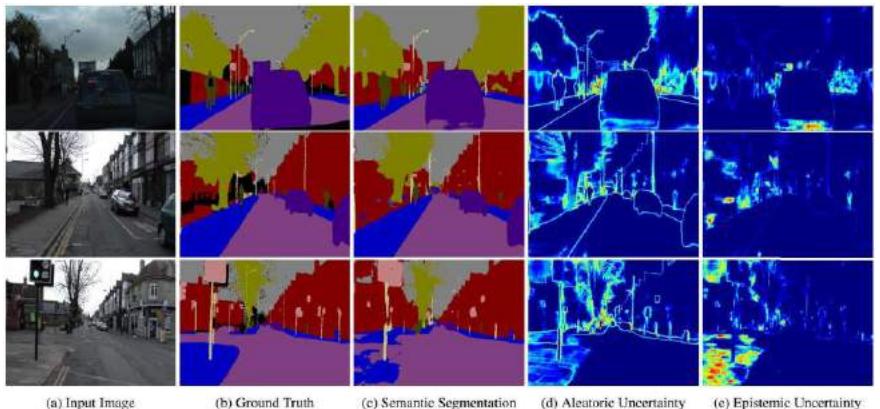
Overview of Explanation in Machine Learning (3)

● Computer Vision



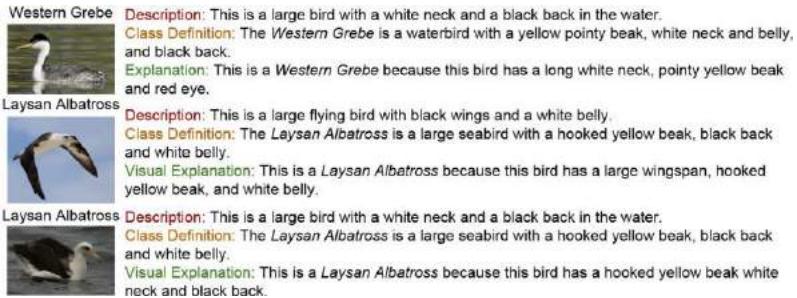
Interpretable Units

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



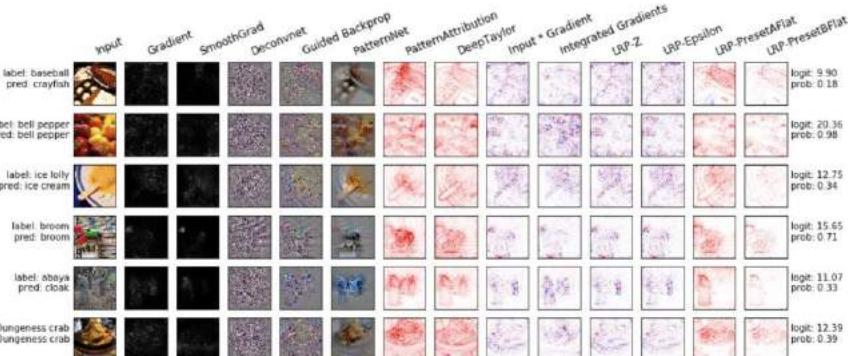
Uncertainty Map

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



Visual Explanation

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

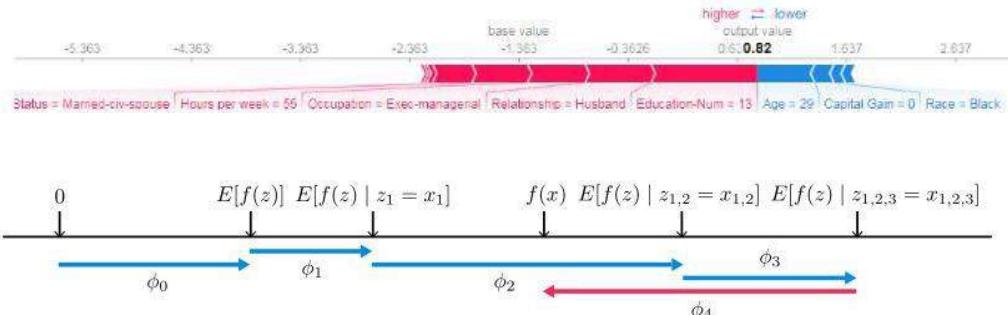


Saliency Map

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

Overview of Explanation in Different AI Fields (1)

• Game Theory

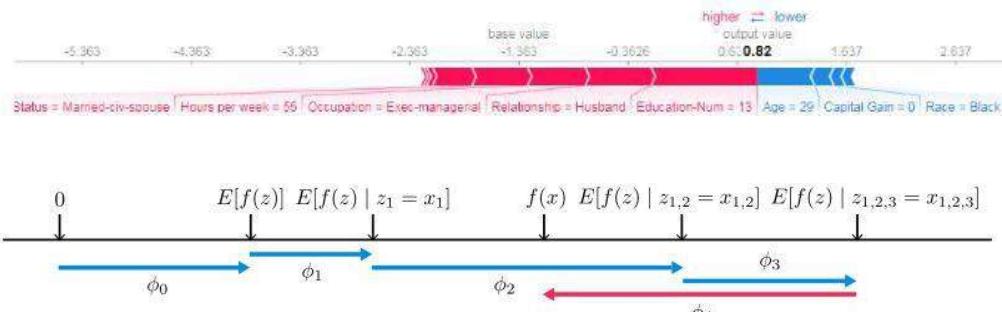


Shapley Additive Explanation

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

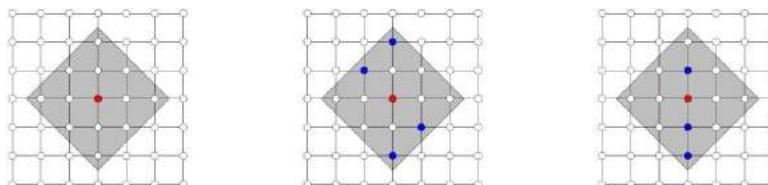
Overview of Explanation in Different AI Fields (1)

• Game Theory



Shapley Additive Explanation

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

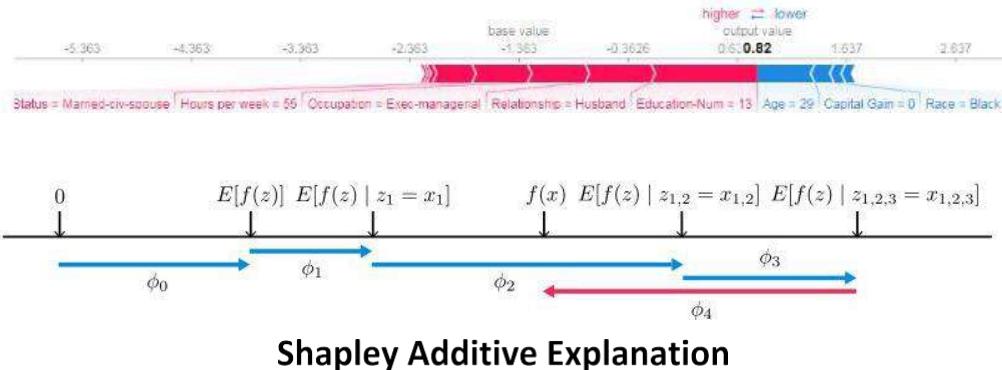


L-Shapley and C-Shapley (with graph structure)

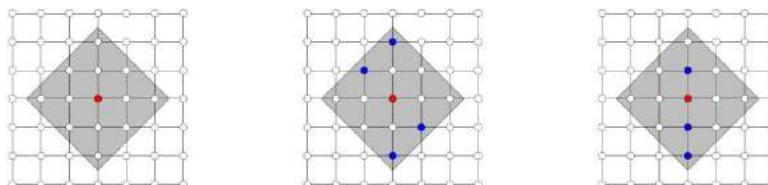
Jianbo Chen, Le Song, Martin J. Wainwright, Michael I. Jordan: L-Shapley and C-Shapley: Efficient Model Interpretation for Structured Data. ICLR 2019

Overview of Explanation in Different AI Fields (1)

• Game Theory



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



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

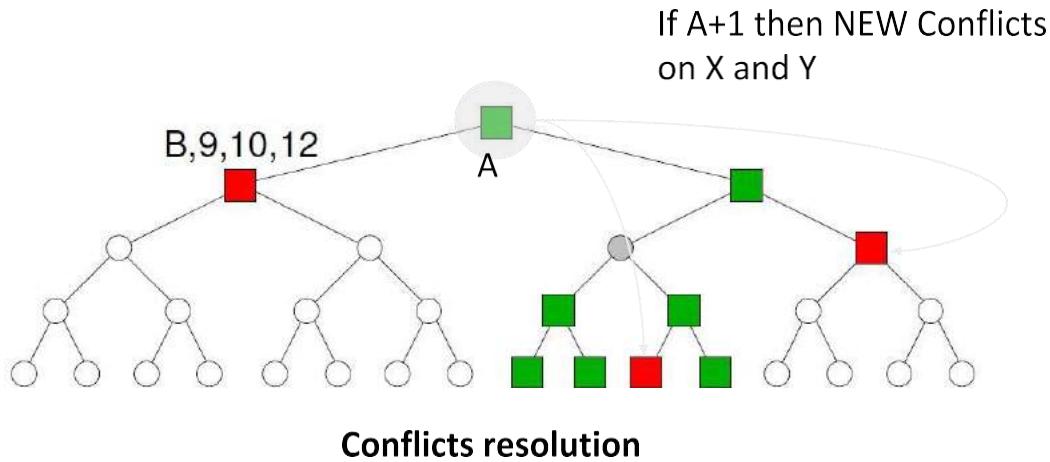
instance-wise feature importance (causal influence)

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



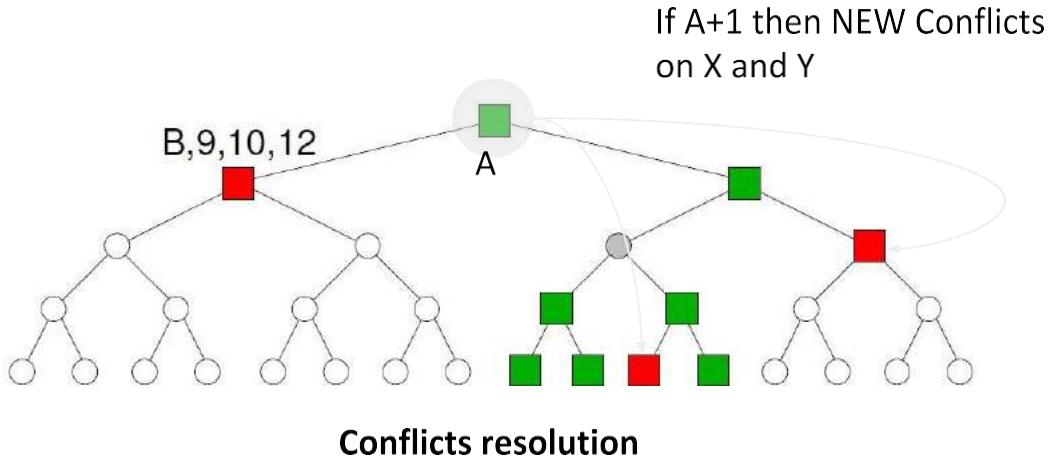
Barry O'Sullivan, Alexandre Papadopoulos, Boi Faltings, Pearl Pu: Representative Explanations for Over-Constrained Problems. AAAI 2007: 323-328

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)

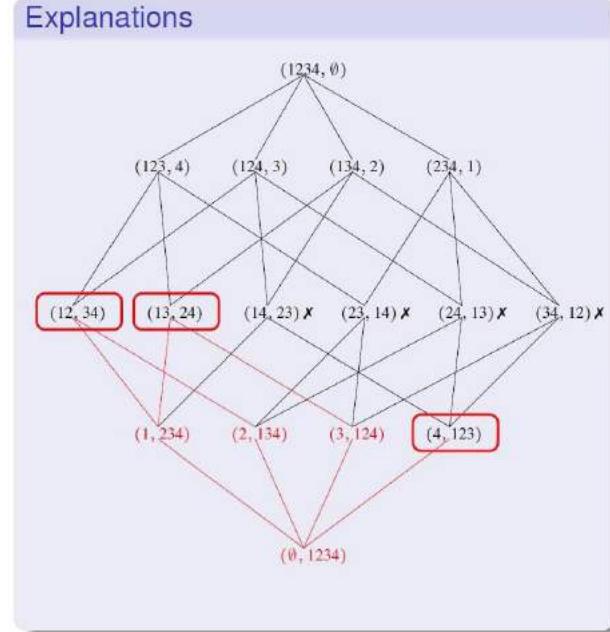
- Search and Constraint Satisfaction



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).



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	$\frac{\vdash c \implies c}{\vdash c \implies c}$	
Trans	$\frac{\vdash c \implies d, \vdash d \implies e}{\vdash c \implies e}$	
Eq	$\frac{\vdash A \equiv B}{\vdash C[A/B] \implies D[A/B]}$	
Prim	$\frac{FF \subseteq BE}{\vdash (\text{prim } BE) \implies (\text{prim } FF)}$	
THING	$\vdash C \implies \text{THING}$	
AndR	$\frac{\vdash c \implies p, \vdash c \implies (\text{and } BE)}{\vdash c \implies (\text{and } p BE)}$	
AndL	$\frac{\vdash c \implies B}{\vdash (\text{and } ... c ...) \implies B}$	
All	$\frac{\vdash c \implies p}{\vdash (\text{all } p c) \implies (\text{all } p B)}$	
AtLst	$\frac{}{\vdash (\text{at-least } n p) \implies (\text{at-least } m p)}$	$n > m$
AndEq	$\vdash C \equiv (\text{and } C)$	
AtLst	$\vdash (\text{at - least } 0 p) \equiv \text{THING}$	
All-thing	$\vdash (\text{all } p \text{ THING}) \equiv \text{THING}$	
All-and	$\vdash (\text{and } (\text{all } p \text{ C }) (\text{all } p \text{ D }) ...) \equiv (\text{and } (\text{all } p (\text{and } C D)) ...)$	$A \equiv (\text{and } (\text{at-least } 3 \text{ grape}) (\text{prim } \text{GOOD WINE}))$

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

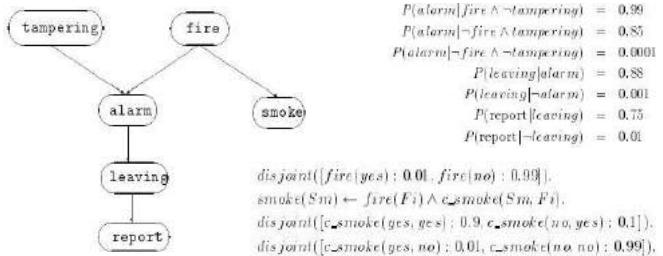
Overview of Explanation in Different AI Fields (3)

- Knowledge Representation and Reasoning

Ref	$\vdash C \Rightarrow C$
Trans	$\vdash c \Rightarrow d, \vdash d \Rightarrow e \vdash c \Rightarrow e$
Eq	$\vdash A \equiv B \vdash c \Rightarrow d \vdash c[A/B] \Rightarrow d[A/B]$
Prim	$\vdash (prim \; E) \Rightarrow (prim \; F)$ $\vdash C \Rightarrow \text{THING}$
THING	$\vdash C \Rightarrow \text{THING}$
AndR	$\vdash c \Rightarrow d, \vdash c \Rightarrow (\text{and } E) \vdash c \Rightarrow (\text{and } d \; E)$
AndL	$\vdash c \Rightarrow b \vdash (\text{and } \dots c \dots) \Rightarrow b$
All	$\vdash c \Rightarrow d \vdash (\text{all } p \; c) \Rightarrow (\text{all } p \; d)$
AtLst	$\vdash (at\text{-}least \; n \; p) \Rightarrow (at\text{-}least \; m \; p)$
AndEq	$\vdash C \equiv (\text{and } C)$
AtL0	$\vdash (\text{at} - \text{least } 0 \; p) \equiv \text{THING}$
All-thing	$\vdash (\text{all } p \; \text{THING}) \equiv \text{THING}$
All-and	$\vdash (\text{and} (\text{all } p \; C) (\text{all } p \; D) \dots) \equiv (\text{and} (\text{all } p \; (\text{and } C \; D)) \dots)$

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

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



Abduction Reasoning (in Bayesian Network)

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

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

Overview of Explanation in Different AI Fields (3)

- Knowledge Representation and Reasoning

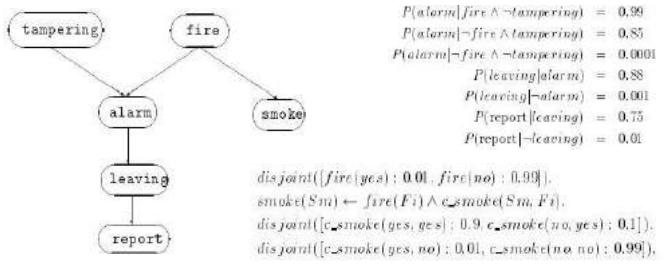
Ref	$\vdash C \Rightarrow C$
Trans	$\vdash c \Rightarrow d, \vdash d \Rightarrow e \vdash c \Rightarrow e$
Eq	$\vdash A \equiv B \vdash c[A/B] \Rightarrow d[A/B]$
Prim	$\vdash (prim \; FF) \Rightarrow (prim \; FF)$
THING	$\vdash C \Rightarrow \text{THING}$
AndR	$\vdash c \Rightarrow d, \vdash c \Rightarrow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

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

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

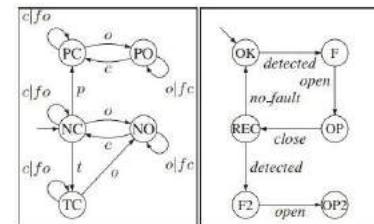
Explaining Reasoning (through Justification) e.g., Subsumption

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



Abduction Reasoning (in Bayesian Network)

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



Diagnosis Inference

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

Overview of Explanation in Different AI Fields (4)

- Multi-agent Systems

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

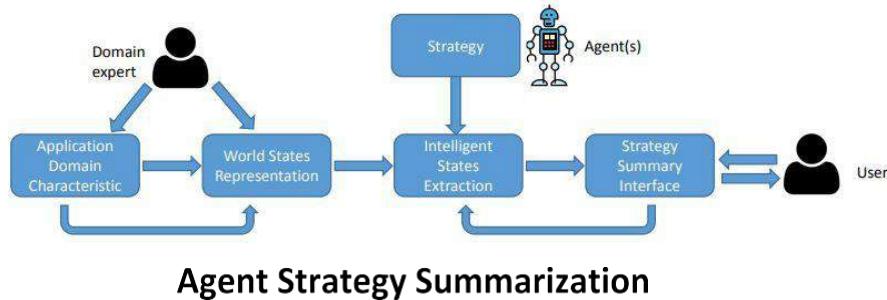
Explanation of Agent Conflicts & Harmful Interactions

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

Overview of Explanation in Different AI Fields (4)

• Multi-agent Systems

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



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

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)

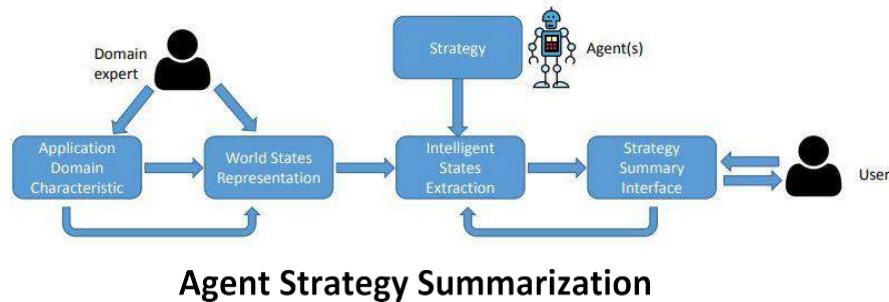
Overview of Explanation in Different AI Fields (4)

• Multi-agent Systems

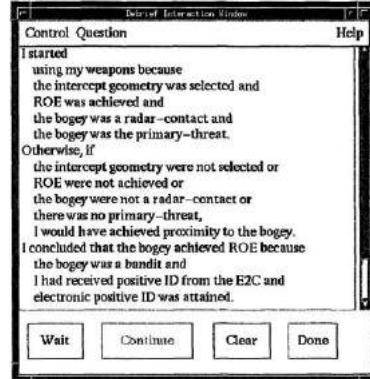
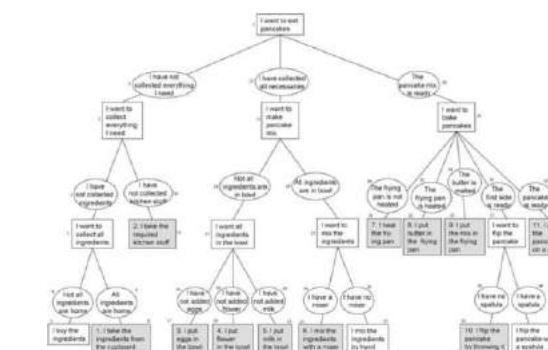
MAS INFRASTRUCTURE		INDIVIDUAL AGENT INFRASTRUCTURE	
MAS INTEROPERATION	Translation Services Interoperation Services	INTEROPERATION	Interoperation Modules
CAPABILITY TO AGENT MAPPING	Middle Agents	CAPABILITY TO AGENT MAPPING	Middle Agents Components
NAME TO LOCATION MAPPING	ANS	NAME TO LOCATION MAPPING	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, Activity Visualization, Launching	MANAGEMENT SERVICES	Logging and Visualization Components
ACL INFRASTRUCTURE	Public Ontology Protocols Servers	ACL INFRASTRUCTURE	ACL Parser Private Ontology Protocol Engine
COMMUNICATION INFRASTRUCTURE	Discovery Message Transfer	COMMUNICATION MODULES	Discovery Component Message Transfer Module
OPERATING ENVIRONMENT		I want to eat pancakes.	The geometry was selected.
Machines, OS, Network		I have collected all the necessary ingredients.	The geometry was selected.
Multicast Transport Layer: TCP/IP, Wireless, Infrared, SSL		I want to make pancakes.	The geometry was selected.

Explanation of Agent Conflicts & Harmful Interactions

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



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



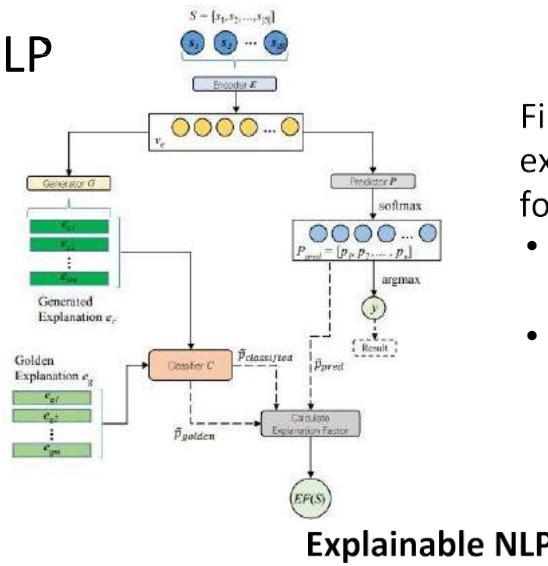
Explainable Agents

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

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

Overview of Explanation in Different AI Fields (5)

- NLP



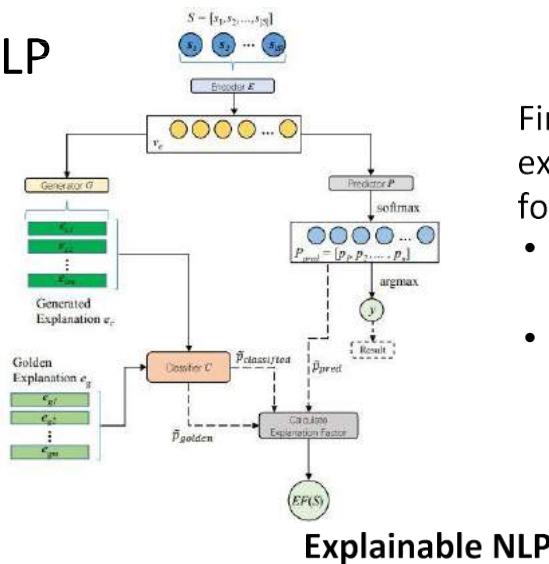
Fine-grained explanations are in the form of:

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

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

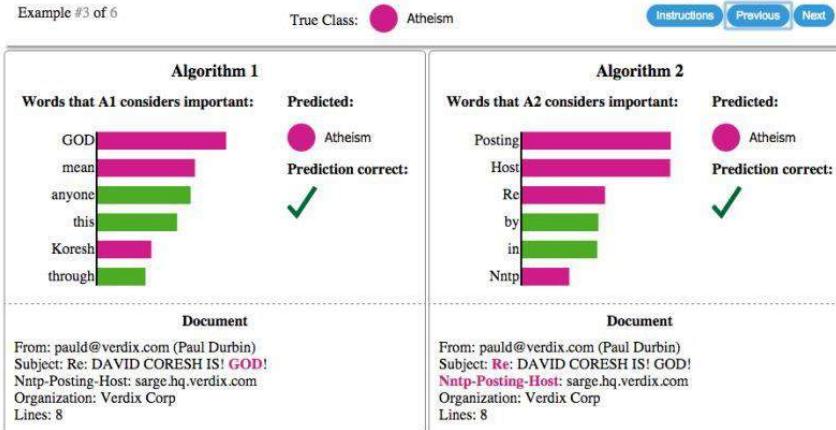
Overview of Explanation in Different AI Fields (5)

• NLP



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

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

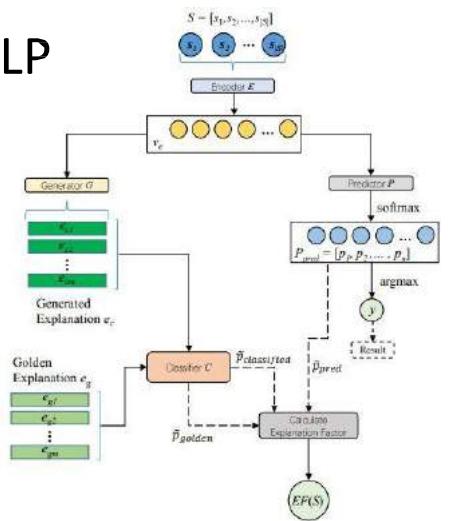


LIME for NLP

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

Overview of Explanation in Different AI Fields (5)

- NLP



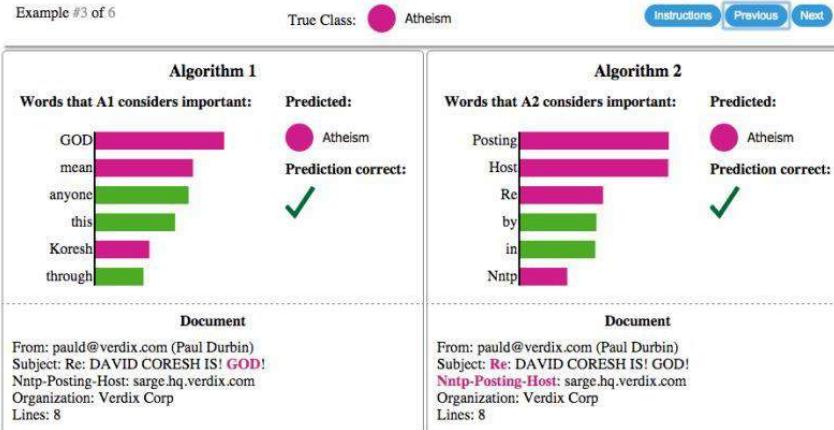
Explainable NLP

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

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

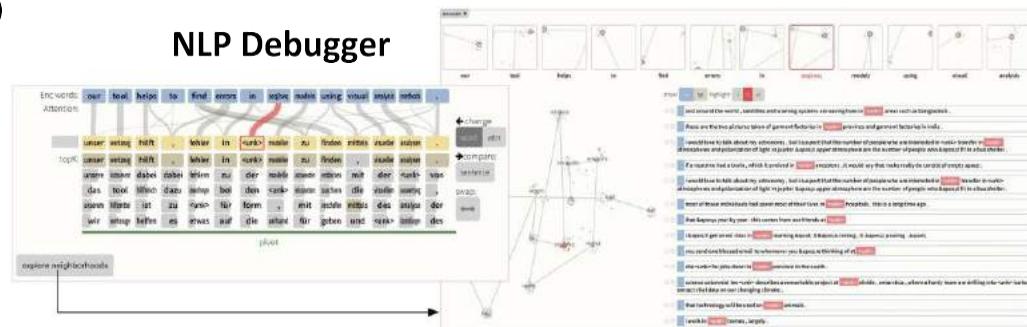
Hendrik Strobelt, Sebastian Gehrmann, Michael Behrisch, Adam Perer, Hanspeter Pfister, Alexander M. Rush: Seq2seq-Vis: A Visual Debugging Tool for Sequence-to-Sequence Models. IEEE Trans. Vis. Comput. Graph. 25(1): 353-363 (2019)

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



LIME for NLP

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

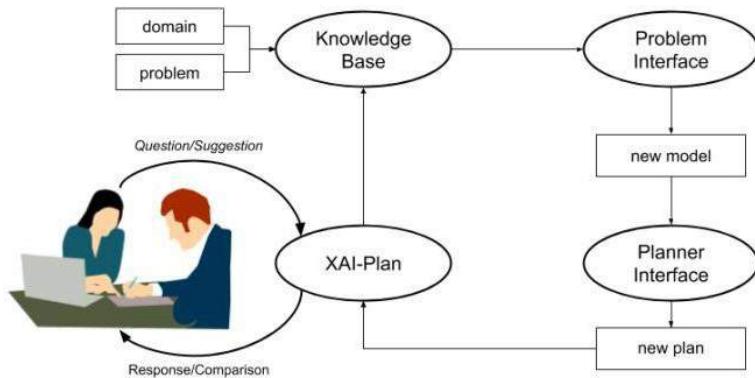


Overview of Explanation in Different AI Fields (6)

- Planning and Scheduling

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

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



XAI Plan

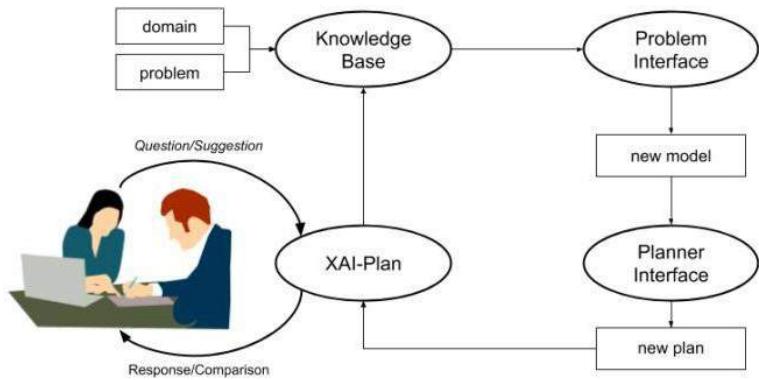
Rita Borgo, Michael Cashmore, Daniele Magazzeni: Towards Providing Explanations for AI Planner Decisions. CoRR abs/1810.06338 (2018)

Overview of Explanation in Different AI Fields (6)

• Planning and Scheduling

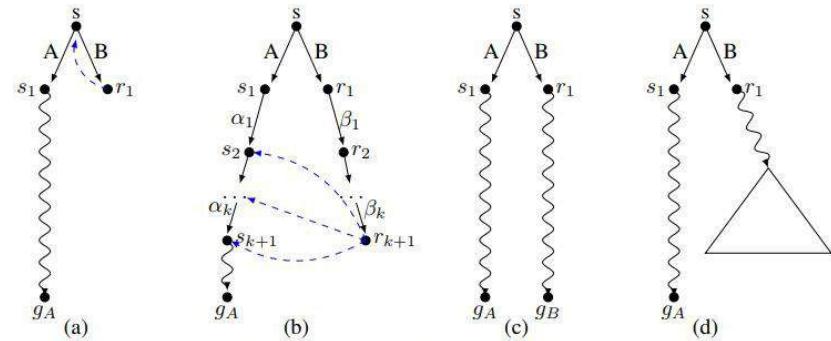
Explanation Type	R1	R2	R3	R4
Plan Patch Explanation / VAL	✗	✓	✗	✓
Model Patch Explanation	✓	✗	✓	✓
Minimally Complete Explanation	✓	✓	✗	?
Minimally Monotonic Explanation	✓	✓	✓	?
(Approximate) Minimally Complete Explanation	✗	✓	✗	✓

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



XAI Plan

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



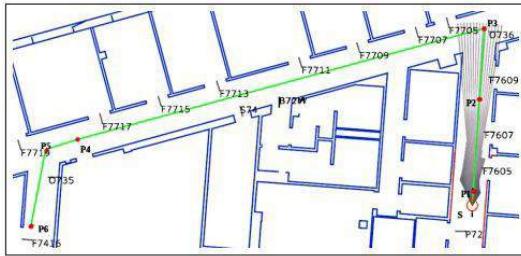
Human-in-the-loop Planning

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

(Manual) Plan Comparison

Overview of Explanation in Different AI Fields (7)

- Robotics



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

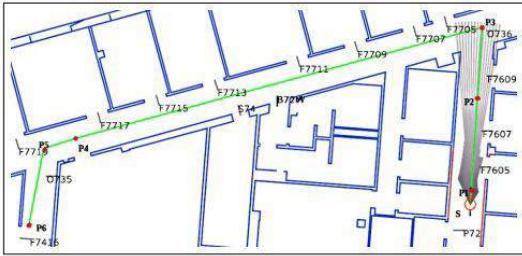
Narration of Autonomous Robot Experience

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

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

Overview of Explanation in Different AI Fields (7)

- Robotics



		Abstraction, A			
		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.

Robot: I have decided to turn left.

Human: Why did you do that?

Robot: I believe that the correct action is to turn left
BECAUSE:

I'm being asked to go forward

AND This area in front of me was 20 cm higher than me
highlights area

AND the area to the left has maximum protrusions of less
than 5 cm *highlights area*

AND I'm tilted to the right by more than 5 degrees.

Here is a display of the path through the tree that lead to
this decision. *displays tree*

Human: How confident are you in this decision?

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

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

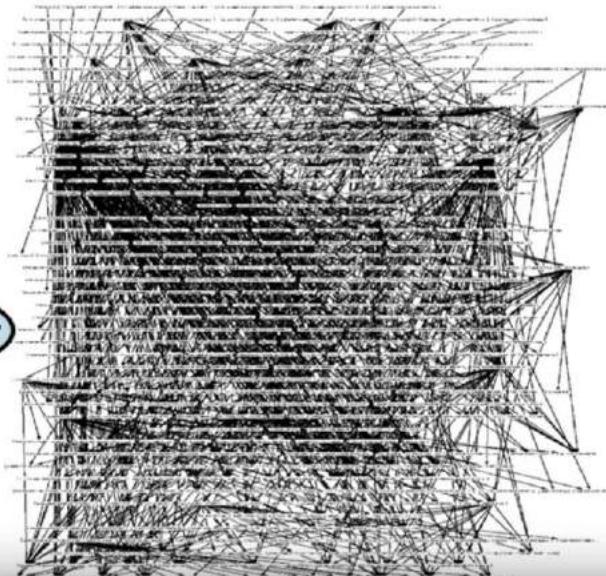
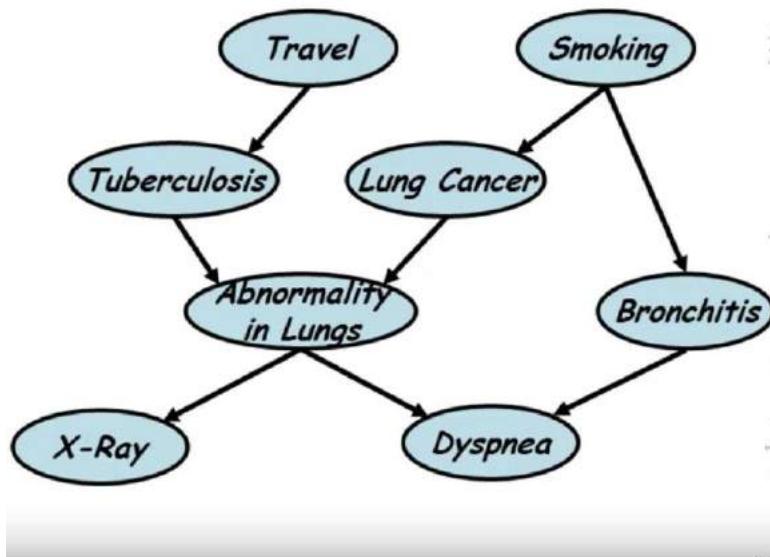
Robot: Here is the histogram of all training examples that
reached this leaf. 80% of examples where this area was
above 20 cm predicted the appropriate action to be “drive
forward”.

From Decision Tree to human-friendly information

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

Overview of Explanation in Different AI Fields (8)

- Reasoning under Uncertainty



Probabilistic Graphical Models

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

Explainable Machine Learning (from a Machine Learning Perspective)

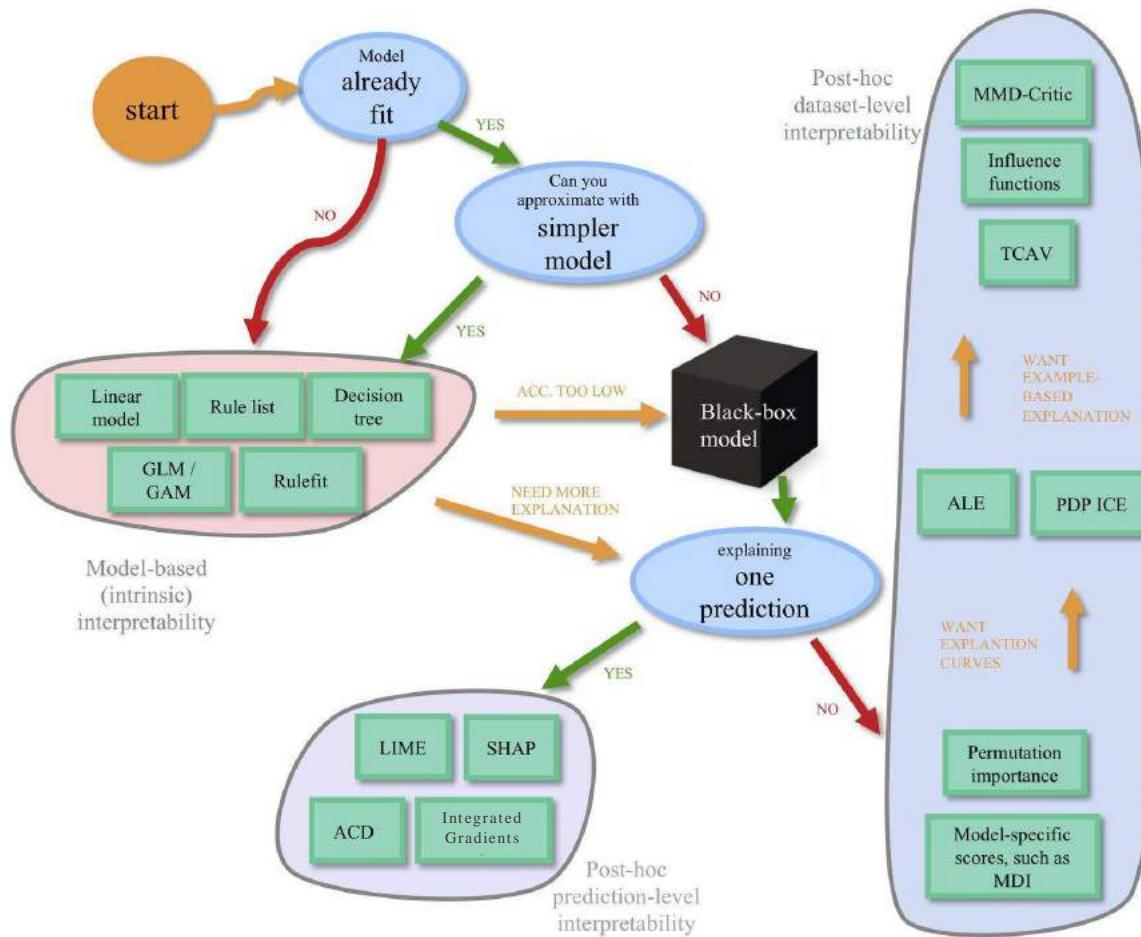
Achieving Explainable AI

Approach 1: Post-hoc explain a given AI model

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

Approach 2: Build an interpretable model

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



interpretability cheat-sheet

[View on github](#)

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

Summaries and links to code

- RuleFit** – automatically add features extracted from a small tree to a linear model
- LIME** – linearly approximate a model at a point
- SHAP** – find relative contributions of features to a prediction
- ACD** – hierarchical feature importances for a DNN prediction
- Text** – DNN generates text to explain a DNN's prediction (sometimes not faithful)
- Permutation importance** – permute a feature and see how it affects the model
- ALE** – perturb feature value of nearby points and see how outputs change
- PDP ICE** – vary feature value of all points and see how outputs change
- TCAV** – see if representations of certain points learned by DNNs are linearly separable
- Influence functions** – find points which highly influence a learned model
- MMD-CRITIC** – find a few points which summarize classes

Achieving Explainable AI

Approach 1: Post-hoc explain a given AI model

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

Approach 2: Build an interpretable model

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



Top label: “**clog**”

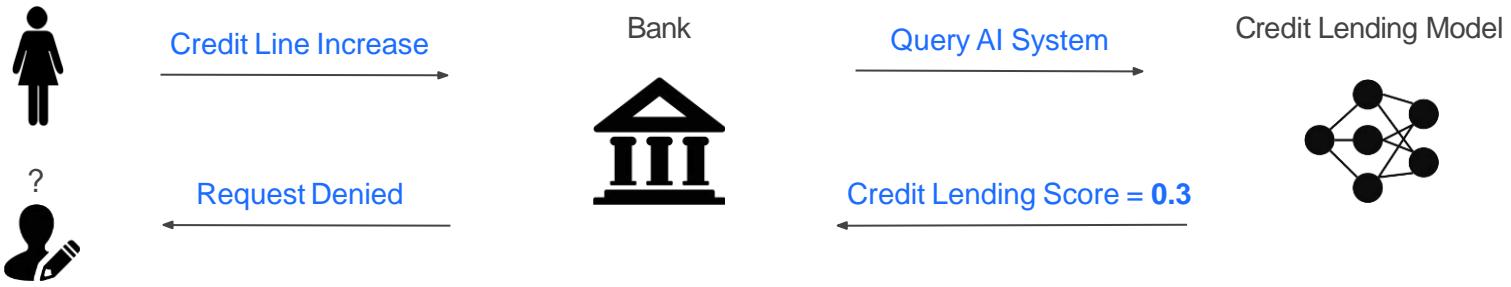
Why did the network label this image as “**clog**”?



Top label: “**fireboat**”

Why did the network label this image as “**fireboat**”?

Credit Lending in a black-box ML world



Why? Why not?

How?

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

The Attribution Problem

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

Examples:

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

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

Application of Attributions

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

Next few slides

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

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

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

**Not a complete list!

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

Ablations

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

Pros:

- Simple and intuitive to interpret

Cons:

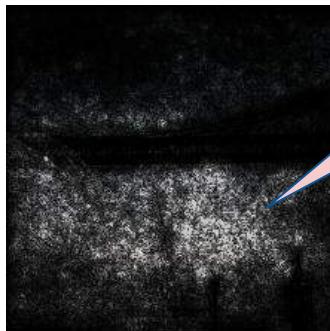
- Unrealistic inputs
- Improper accounting of interactive features
- Can be computationally expensive



Feature*Gradient

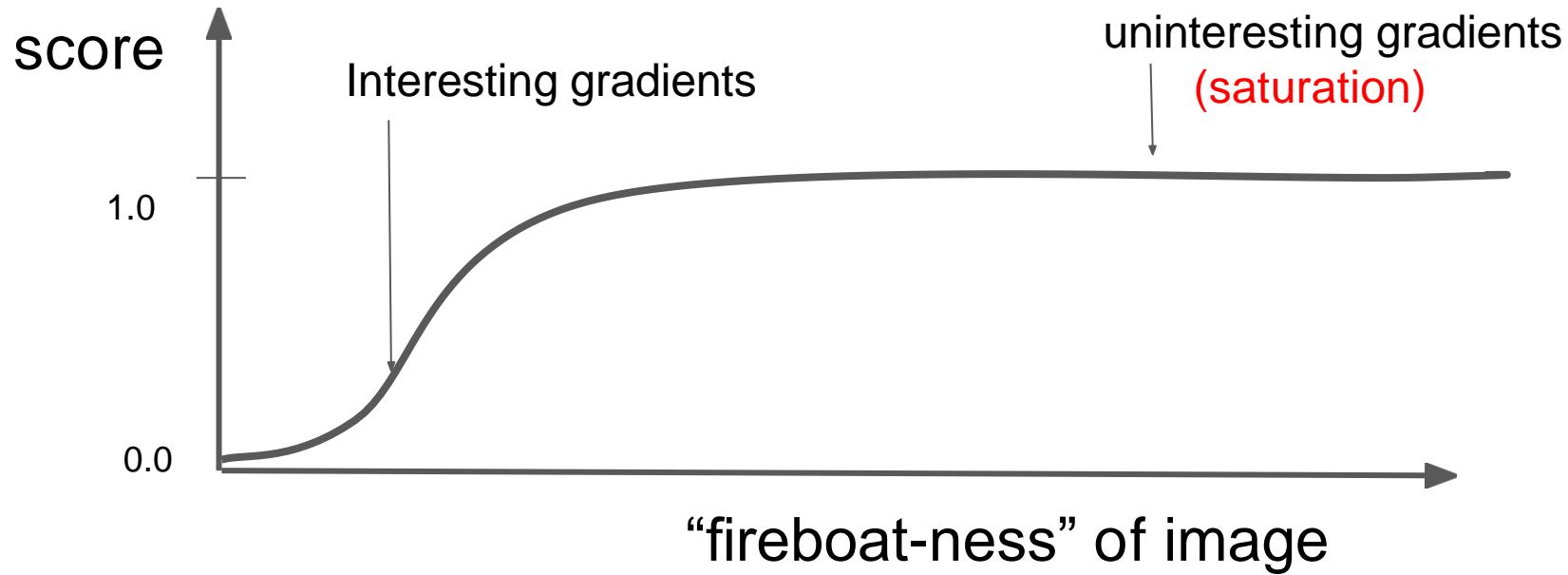
Attribution to a feature is feature value times gradient, i.e., $x_i^* \partial y / \partial x_i$

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



Gradients in the vicinity of the input seem like noise?

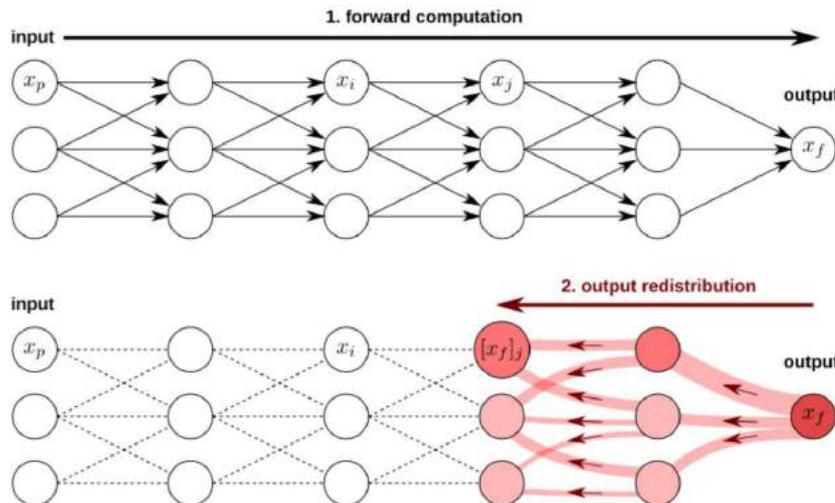
Local linear approximations can be too local



Score Back-Propagation based Methods

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

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



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

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

Score Back-Propagation based Methods

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

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

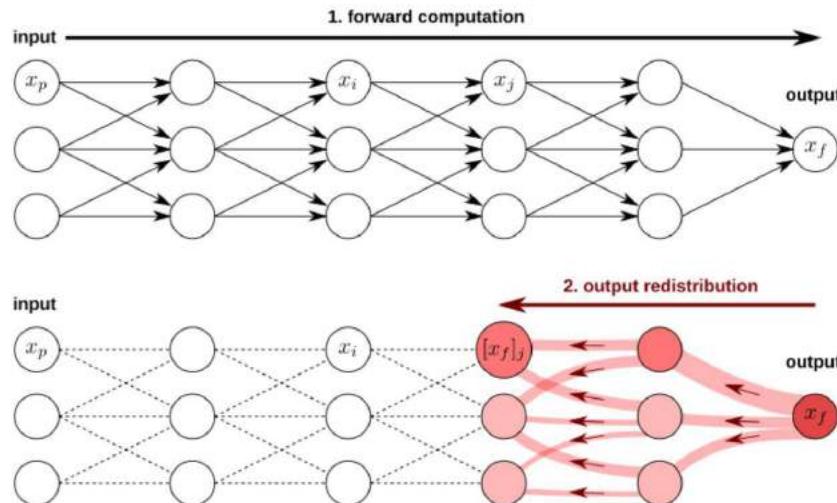


Image credit heatmapping.org

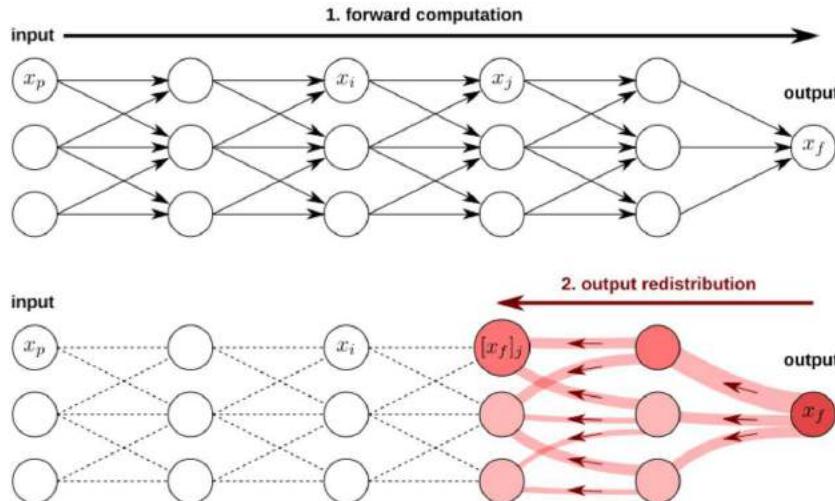
Tricky case: Output of a neuron is a **non-linear** function, e.g., ReLU, Sigmoid, etc.

- **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]



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
 $G(x, y, z) = x * (y * z)$

Image credit heatmapping.org

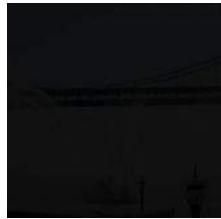
Baselines and additivity

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

Another approach: gradients at many points



Baseline



...scaled inputs ...

...gradients of scaled inputs ...



Input



Integrated Gradients [ICML 2017]

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

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

Original image



Integrated Gradients



Integrated Gradients in action

Why is this image labeled as “clog”?

Original image



“Clog”



Why is this image labeled as “clog”?

Original image



Integrated Gradients
(for label “clog”)

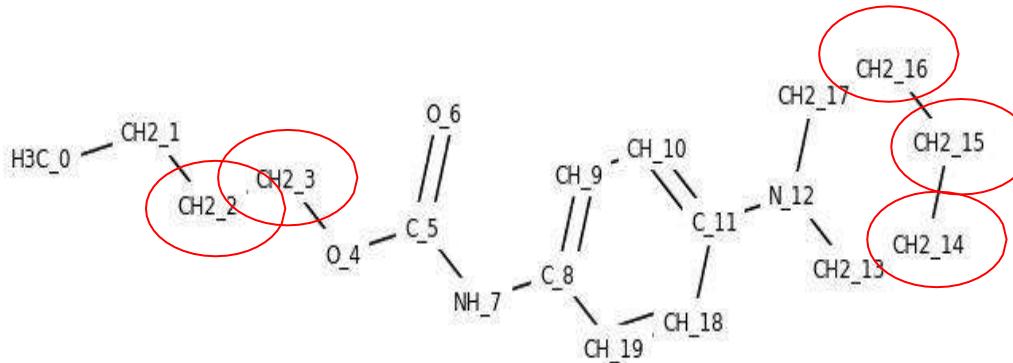


“Clog”



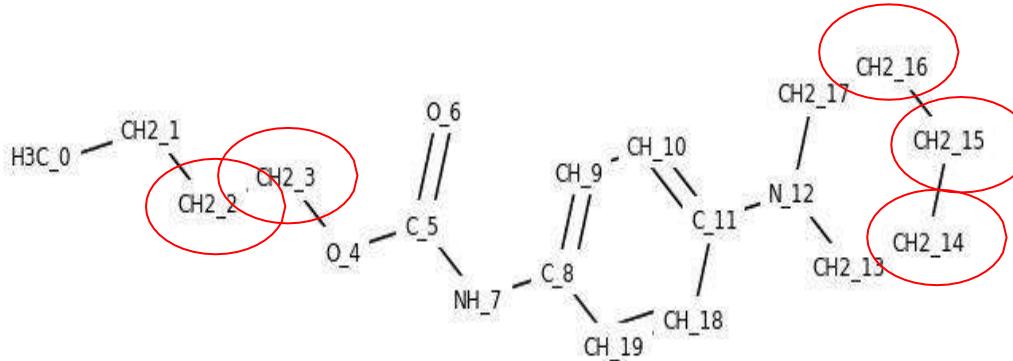
Detecting an architecture bug

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



Detecting an architecture bug

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



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

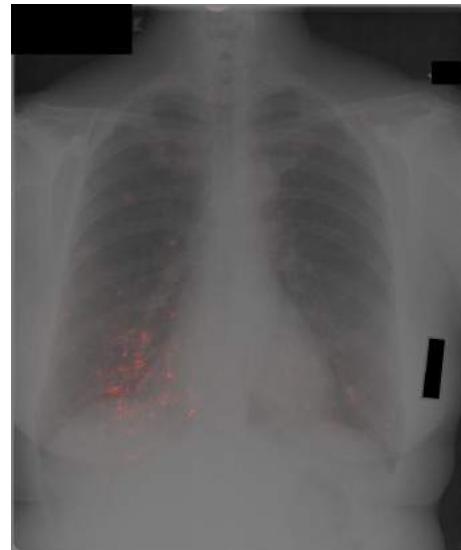
Detecting a data issue

- Deep network predicts various diseases from chest x-rays

Original image



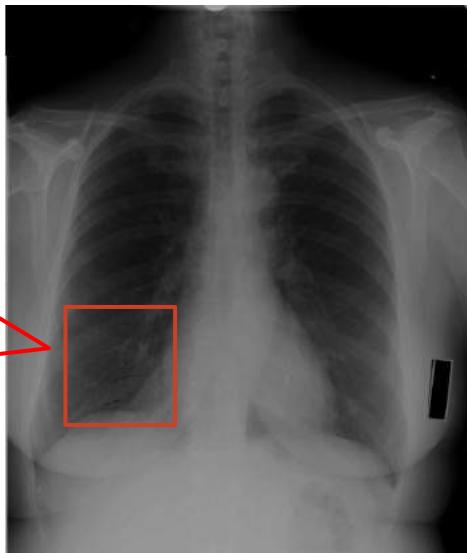
Integrated gradients
(for top label)



Detecting a data issue

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

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

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

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

Shapley Value Justification

Shapley values are unique under four simple axioms

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

Shapley Values for Explaining ML models

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

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

Shapley Values for Explaining ML models

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

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

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

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

Modeling Feature Absence

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

Different approaches choose different distributions

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

Computing Shapley Values

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

- Shapley values can be expressed as an expectation of marginals

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

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

Non-atomic Games: Aumann-Shapley Values and IG

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

Lessons learned: baselines are important

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

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

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

Some limitations and caveats for attributions

Attributions don't explain everything

Some things that are missing:

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

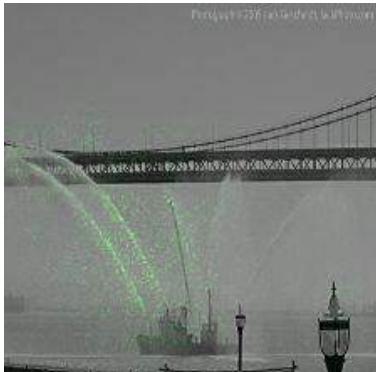
An instance where attributions are useless:

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

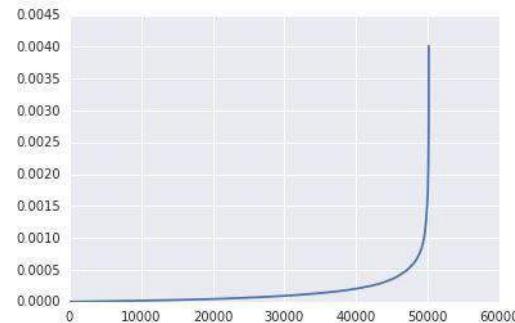
Attributions are for human consumption

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

Naive scaling of attributions
from 0 to 255



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



After clipping attributions
at 99% to reduce range



Other individual prediction explanation methods

Local Interpretable Model-agnostic Explanations

(Ribeiro et al. KDD 2016)

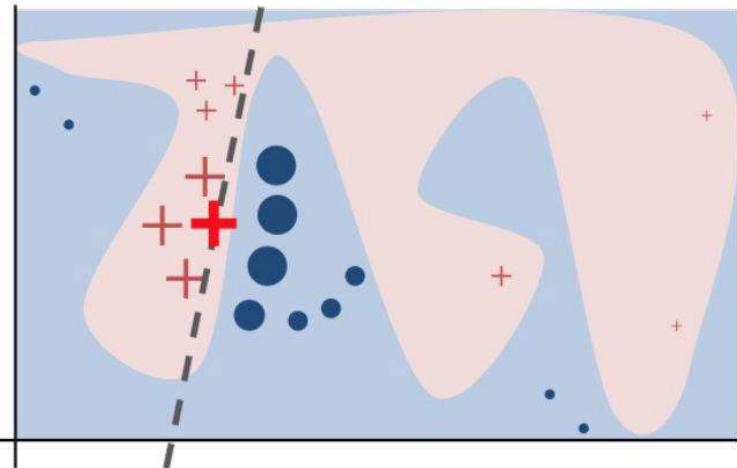
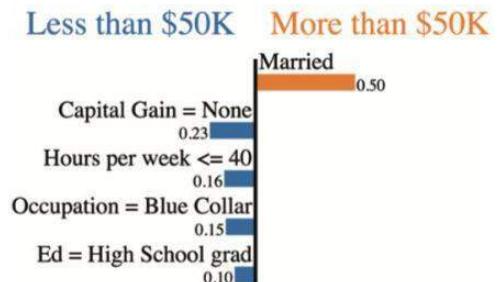


Figure credit: Ribeiro et al. KDD 2016

$28 < \text{Age} \leq 37$
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

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

(a) Instance and prediction



(b) LIME explanation

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

Anchors

$28 < \text{Age} \leq 37$

Workclass = Private

Education = High School grad

Marital Status = Married

Occupation = Blue-Collar

Relationship = Husband

Race = White

Sex = Male

Capital Gain = None

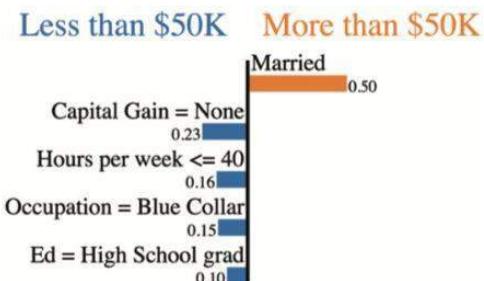
Capital Loss = Low

Hours per week ≤ 40.00

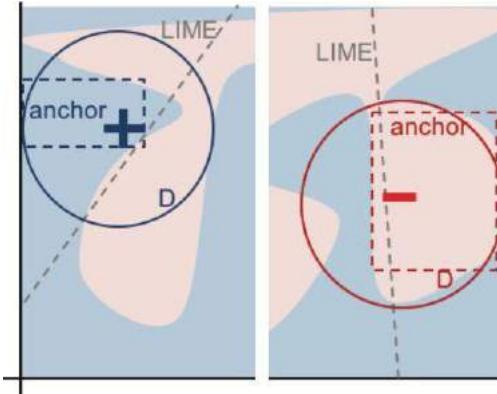
Country = United-States

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

(a) Instance and prediction



(b) LIME explanation



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

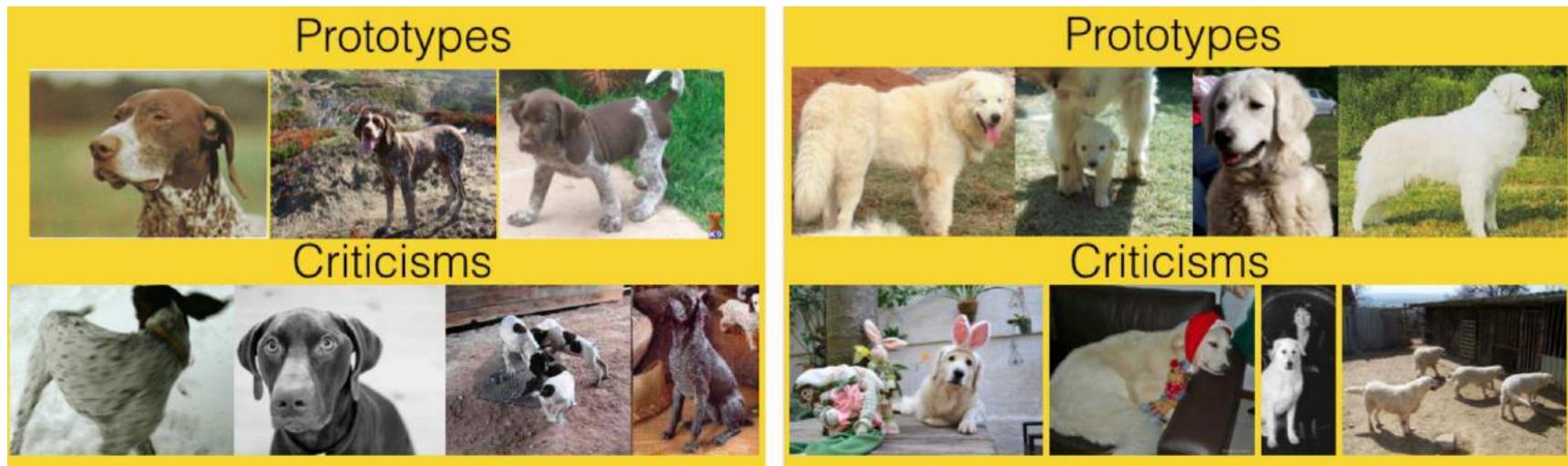
- Trace a model's prediction through the learning algorithm and back to its training data
- Training points “responsible” for a given prediction

Test image



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

Example based Explanations



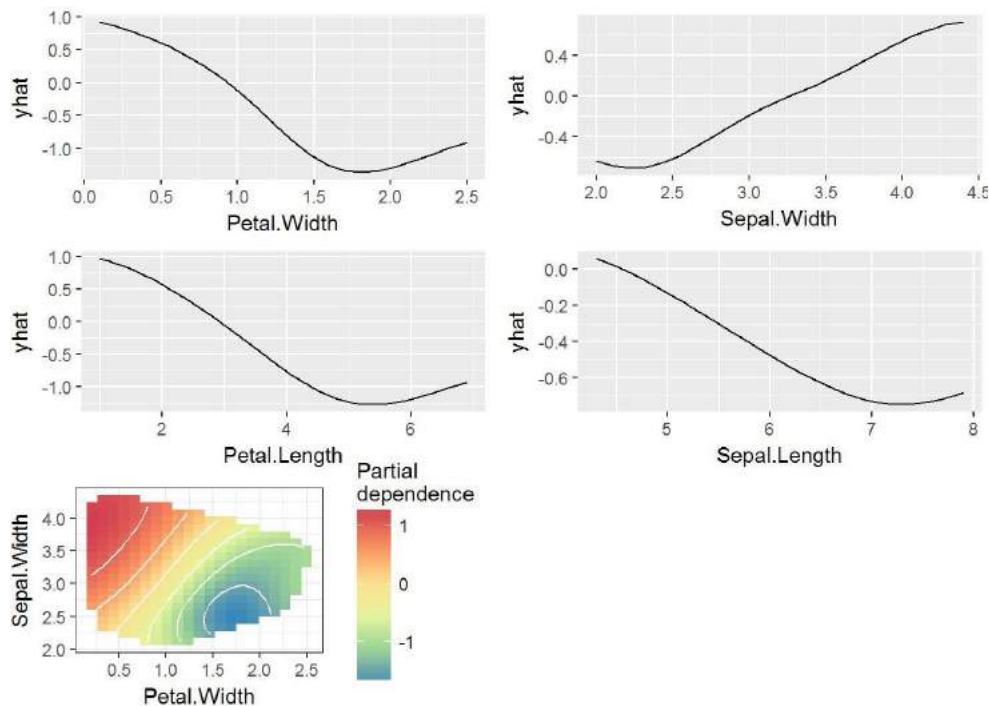
Learned prototypes and criticisms from Imagenet dataset (two types of dog breeds)

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

Global Explanations

Global Explanations Methods

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



Global Explanations Methods

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

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

Random Shuffle of the first feature

Achieving Explainable AI

Approach 1: Post-hoc explain a given AI model

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

Approach 2: Build an interpretable model

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

Decision Trees

Is the person fit?

Age < 30 ?

Yes



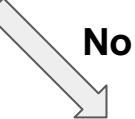
Eats a lot of pizzas?

Yes



Unfit

No



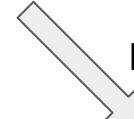
Fit

Exercises in the morning?

Yes



No



Unfit

Optimal Sparse Decision Trees

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

¹Carnegie Mellon University, xiyanghu@cmu.edu

²Duke University, cynthia@cs.duke.edu

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

Decision Set

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

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

..., *then* $Y = 1$. Models of this form have the advantage of being interpretable to human experts

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

Decision List

```
If Past-Respiratory-Illness =Yes and Smoker =Yes and Age ≥ 50, then Lung Cancer  
Else if Allergies =Yes and Past-Respiratory-Illness =Yes, then Asthma  
Else if Family-Risk-Respiratory =Yes, then Asthma  
Else if Family-Risk-Depression =Yes, then Depression  
Else if Gender =Female and Short-Breath-Symptoms =Yes, then Asthma  
Else if BMI ≥ 0.2 and Age≥ 60, then Diabetes  
Else if Frequent-Headaches =Yes and Dizziness =Yes, then Depression  
Else if Frequency-Doctor-Visits ≥ 0.3, then Diabetes  
Else if Disposition-Tiredness =Yes, then Depression  
Else if Chest-Pain =Yes and Nausea and Yes, then Diabetes  
Else Diabetes
```

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

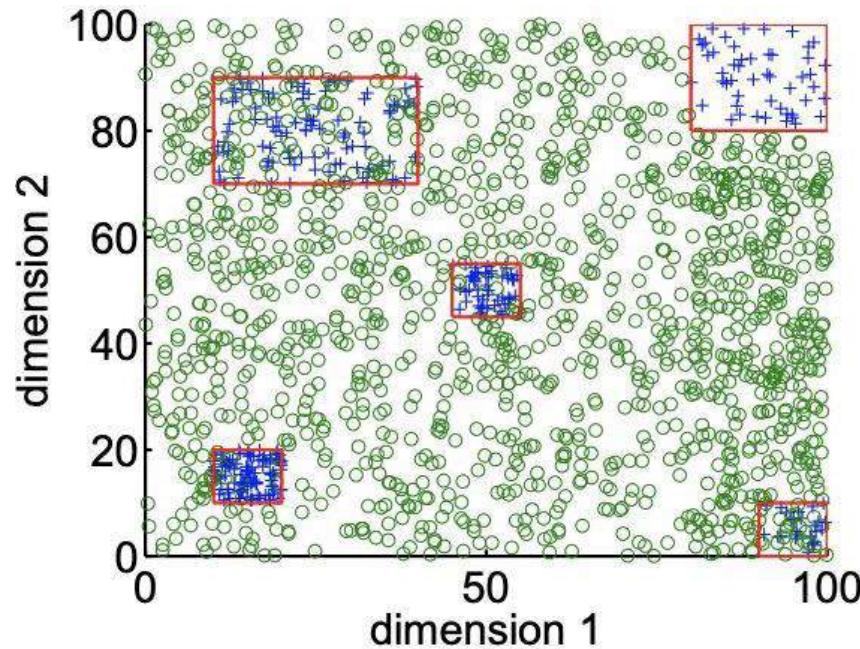


Figure credit: Box Drawings for Learning with Imbalanced. Data Siong Thye Goh and Cynthia Rudin

Supersparse Linear Integer Models for Optimized Medical Scoring Systems

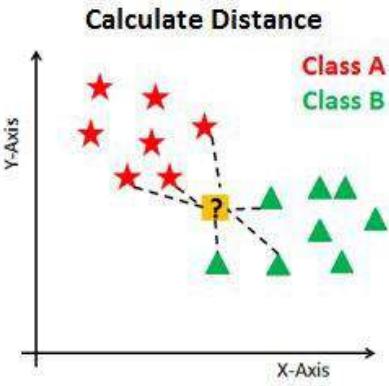
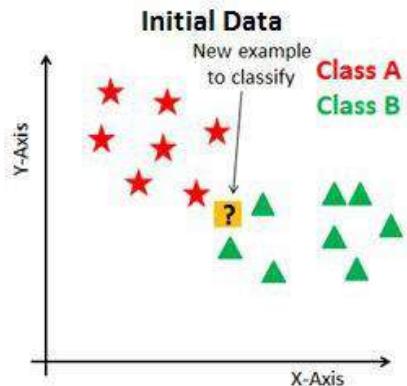
PREDICT PATIENT HAS OBSTRUCTIVE SLEEP APNEA IF SCORE > 1

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

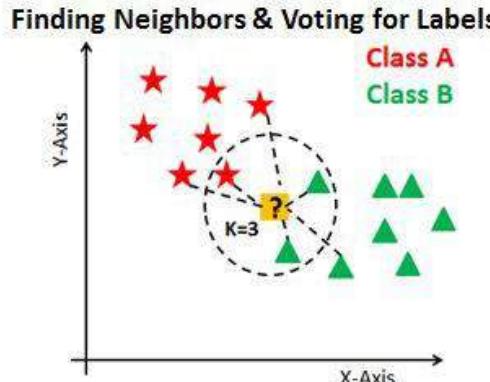
SLIM scoring system for sleep apnea screening. This model achieves a 10-CV mean test TPR/FPR of 61.4/20.9%, obeys all operational constraints, and was trained without parameter tuning. It also generalizes well due to the simplicity of the hypothesis space: here the training TPR/FPR of the final model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

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

K- Nearest Neighbors



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



GLMs and GAMs

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

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

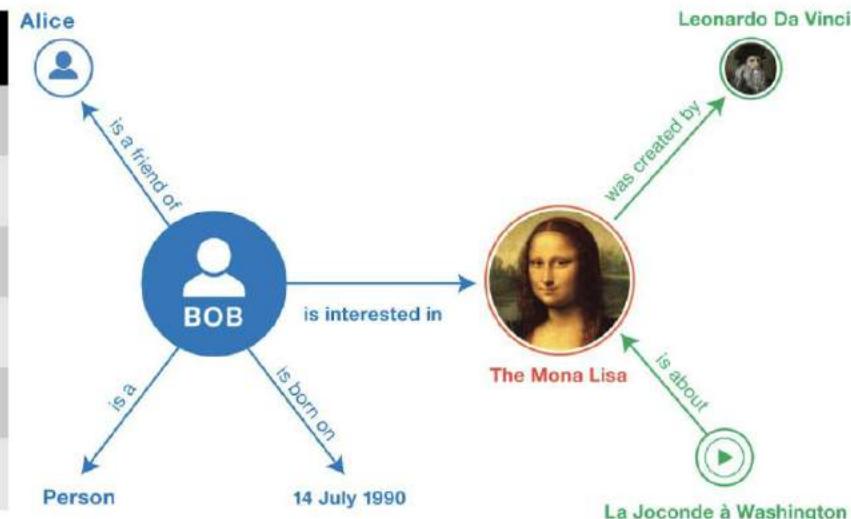
Accurate Intelligible Models with Pairwise Interactions. Lou, Caruana, Gehrke and Hooker. KDD 2013

Explainable Machine Learning (from a Knowledge Graph Perspective)

Knowledge Graph (1)

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

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



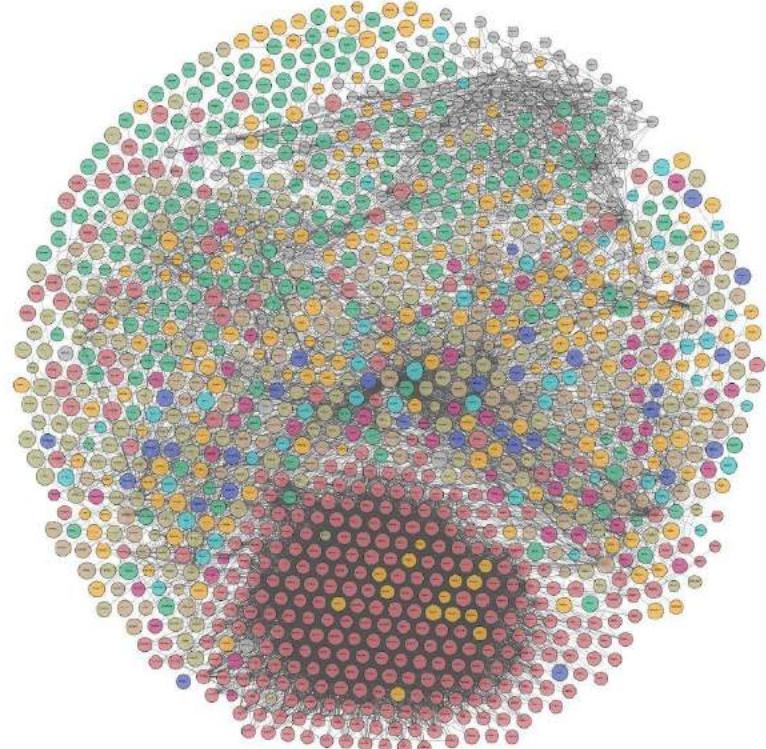
Knowledge Graph (2)

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

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

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

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



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:

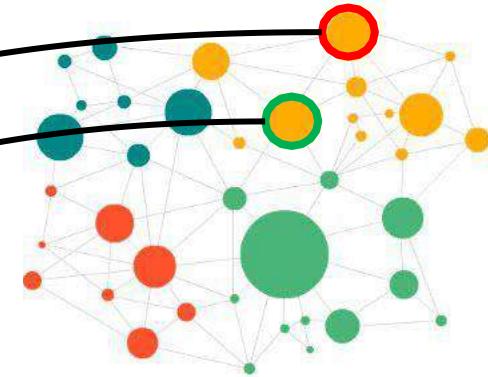
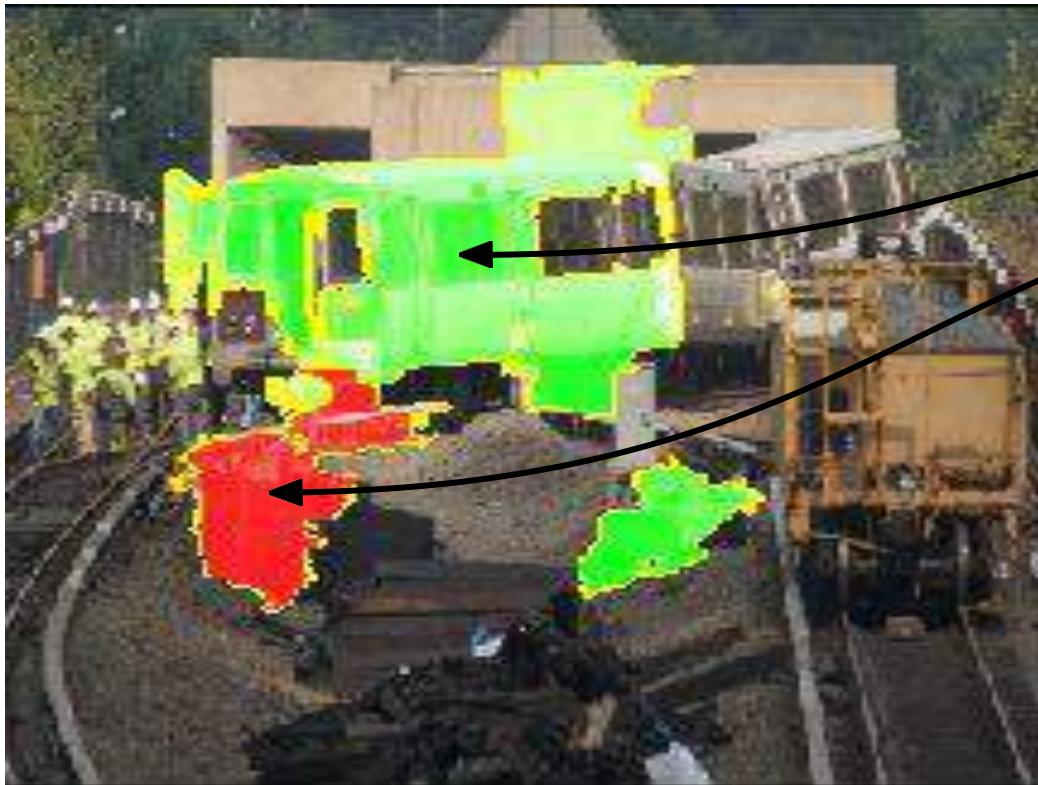
- **Manual** — curated (e.g. via experts), collaborative (e.g. via volunteers)
- **Automated** — semi-structured (e.g. from infoboxes), unstructured (e.g. from text)

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.

Knowledge Graph in Machine Learning (1)

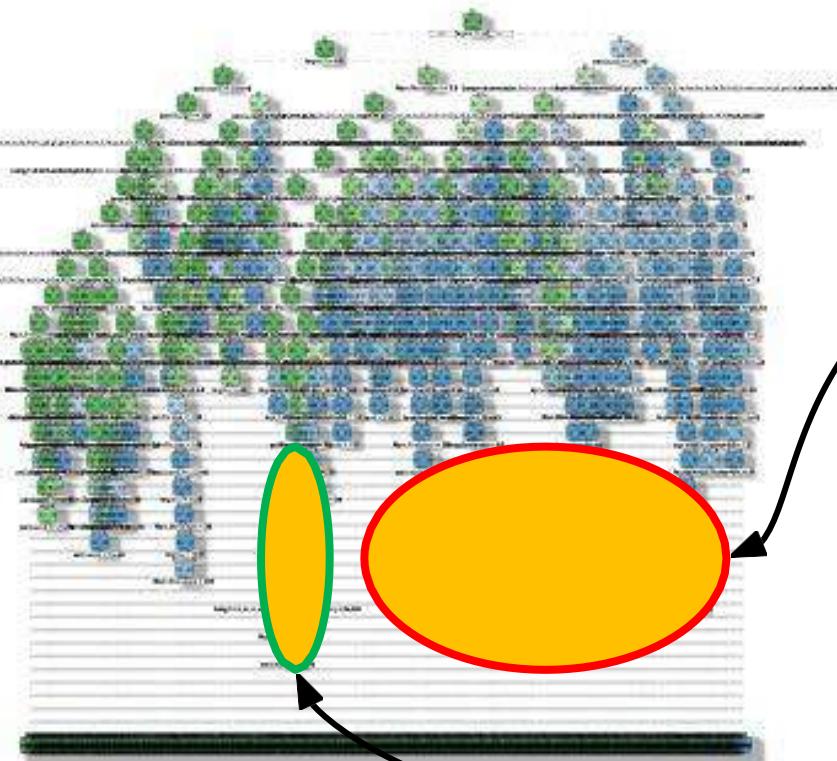


Augmenting (input) features
with more semantics such as
knowledge graph embeddings /
entities

<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)

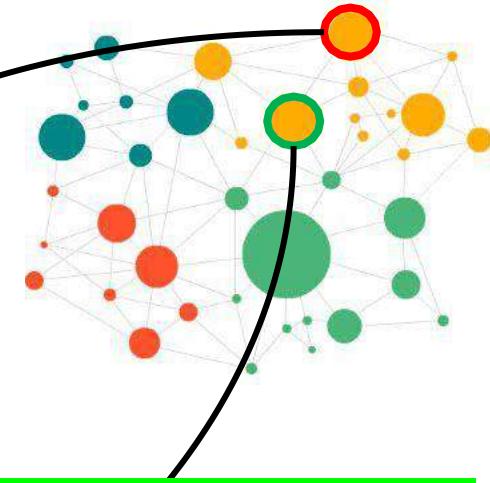
Knowledge Graph in Machine Learning (2)



<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

Training Data

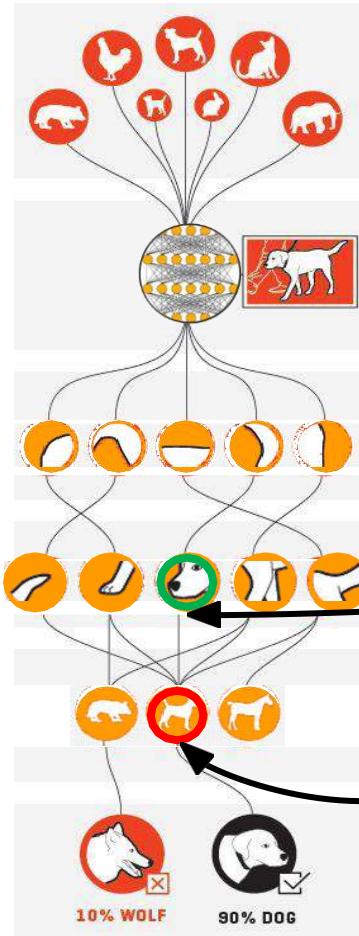
○ Hidden Layer

● Output Layer

Neurons respond to simple shapes

Neurons respond to more complex structures

Neurons respond to highly complex, abstract concepts



Input
(unlabeled
image)

1st Layer

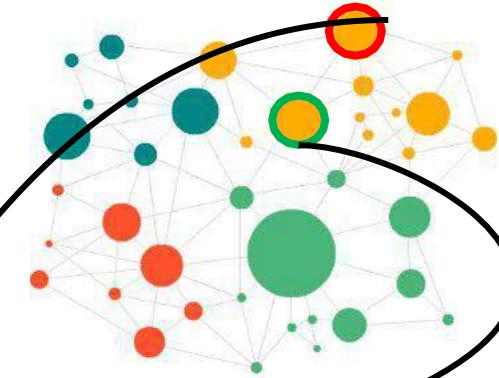
2nd Layer

nth Layer

Low-level
features to
high-level
features

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

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



Knowledge Graph in Machine Learning (4)

● Input Layer

Training Data

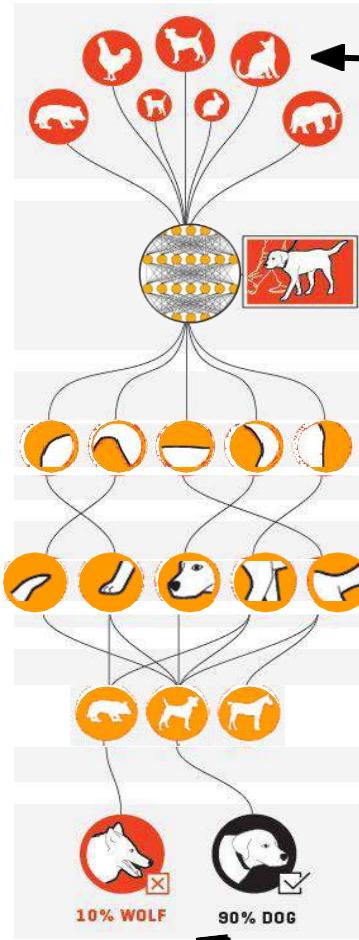
○ Hidden Layer

● Output Layer

Neurons respond to simple shapes

Neurons respond to more complex structures

Neurons respond to highly complex, abstract concepts



Input
(unlabeled
image)

1st Layer

2nd Layer

nth Layer

Low-level
features to
high-level
features

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

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

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

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

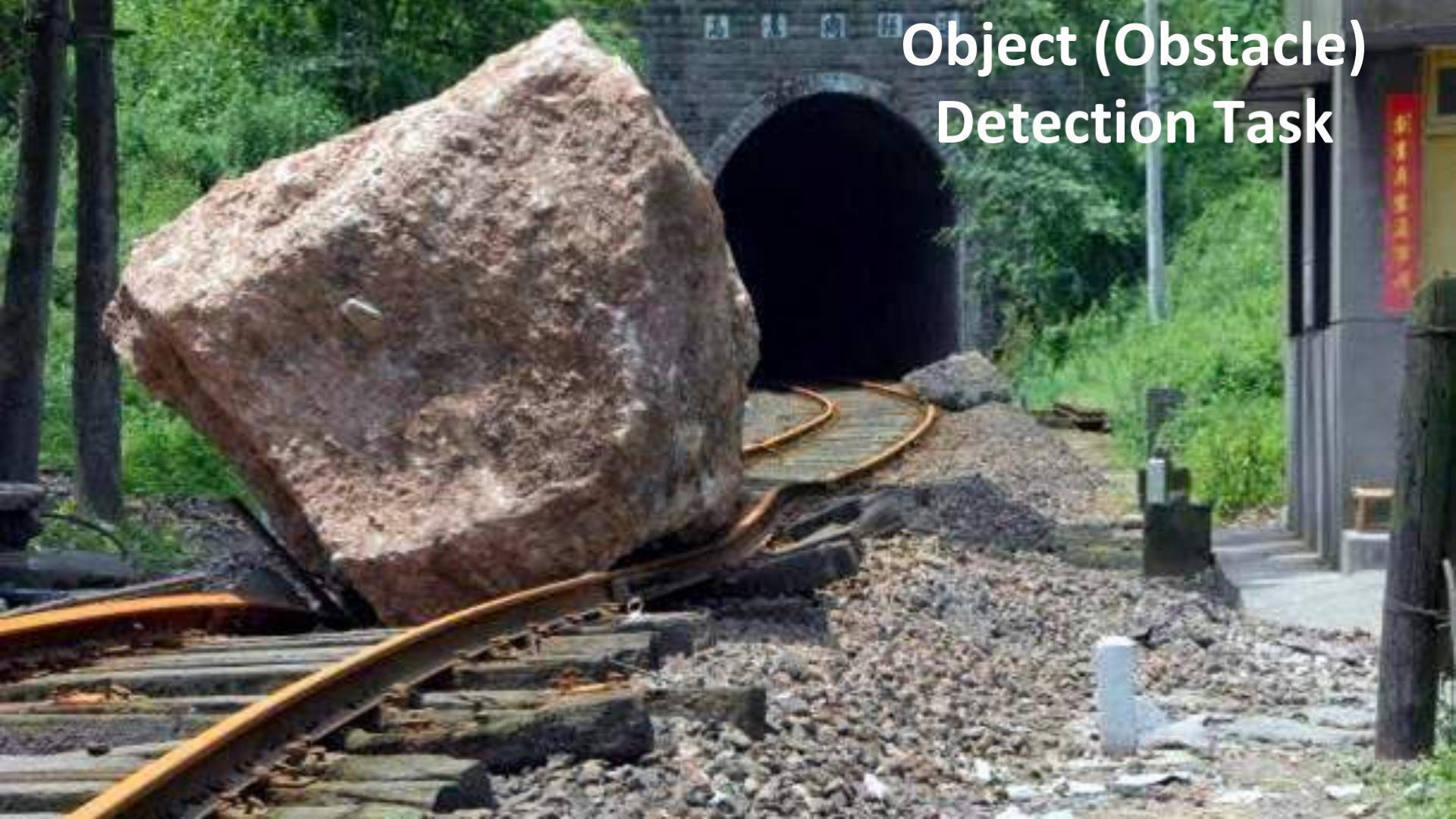
Freddy Lecue
INRIA, France
Accenture Labs, Ireland

Ian Horrocks
Department of Computer Science
University of Oxford, UK

How Does
it
Work
in Practice?

**State of the Art
Machine Learning
Applied to Critical
Systems**

Object (Obstacle) Detection Task

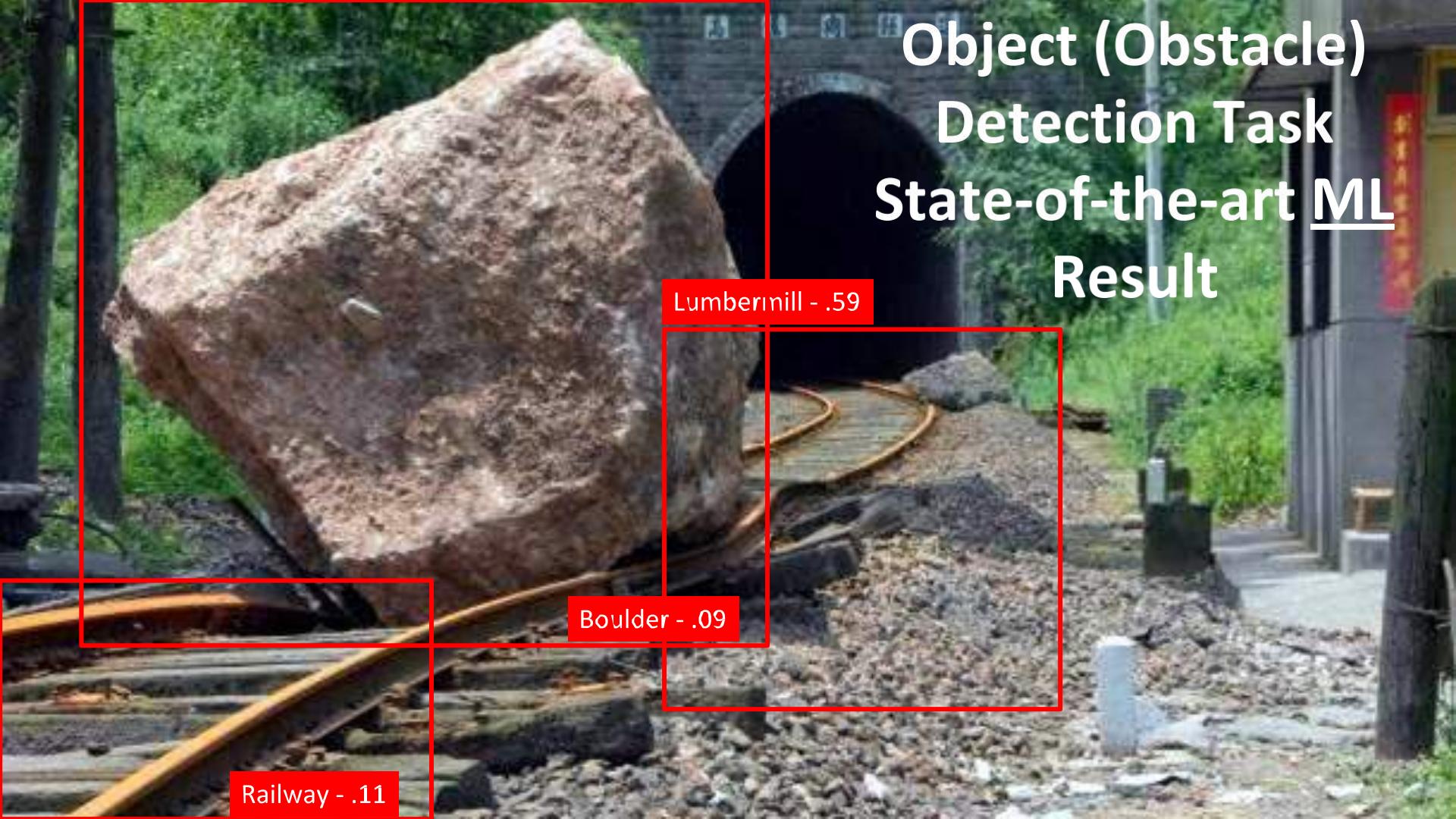


Object (Obstacle) Detection Task State-of-the-art ML Result



Lumbermill - .59

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



State of the Art

XAI

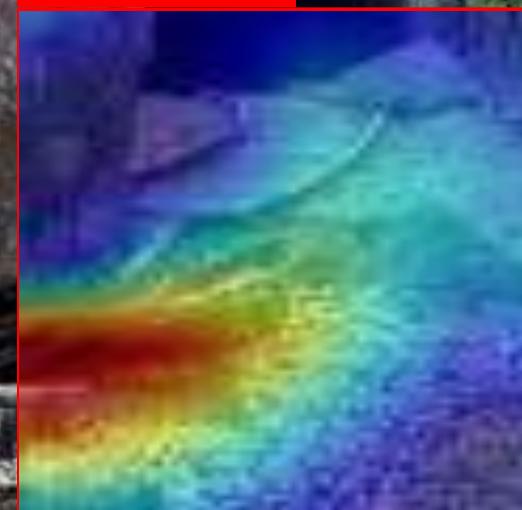
Applied to Critical

Systems

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



Lumbermill - .59



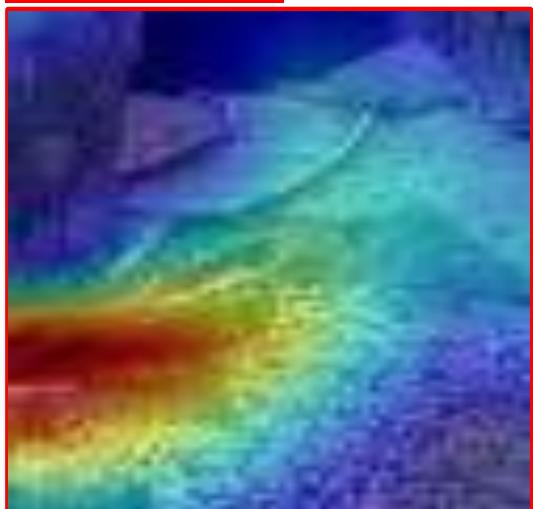
**Unfortunately, this is of
NO use for a human
behind the system**

Let's stay back

Why this Explanation? (meta explanation)

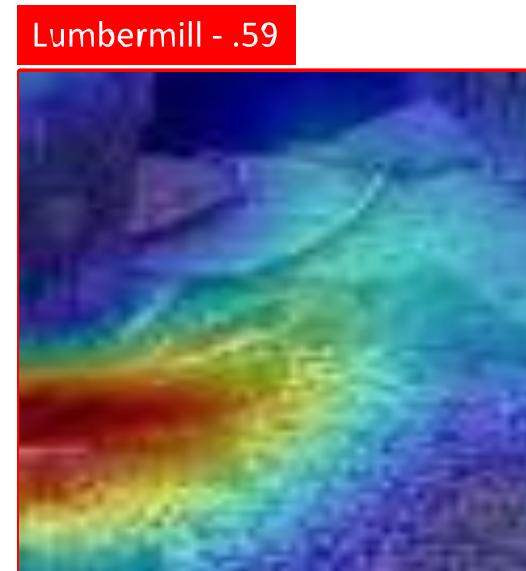
After Human Reasoning...

Lumbermill - .59



DBpedia	
	Browse using ▾
	Formats ▾
dbo:wikiPageID	▪ 352327 (xsd:integer)
dbo:wikiPageRevisionID	▪ 734430894 (xsd:integer)
dc:subject	▪ dbo:Sawmills ▪ dbo:Saws ▪ dbo:Ancient_Roman_technology ▪ dbo:Timber_preparation ▪ dbo: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-es:Sawmill ▪ dbpedia-de:Sawmill ▪ dbpedia-es:Sawmill

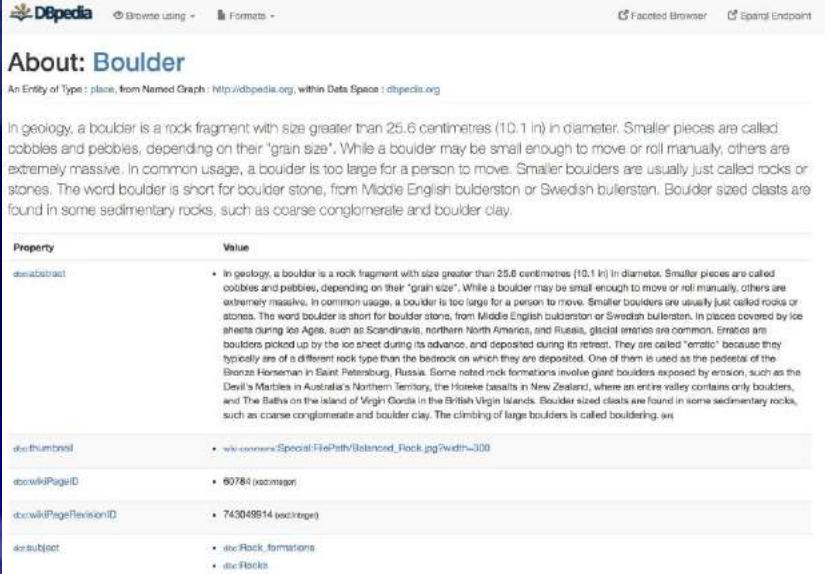
What is missing?



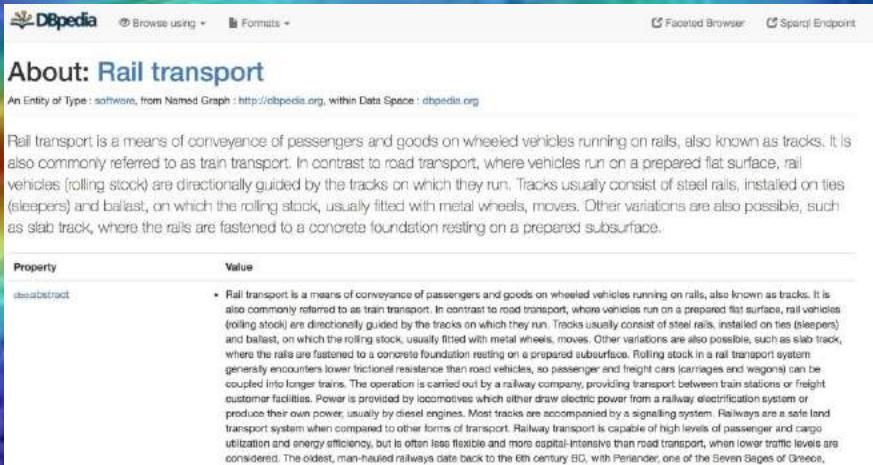
Context matters

Railway - .11

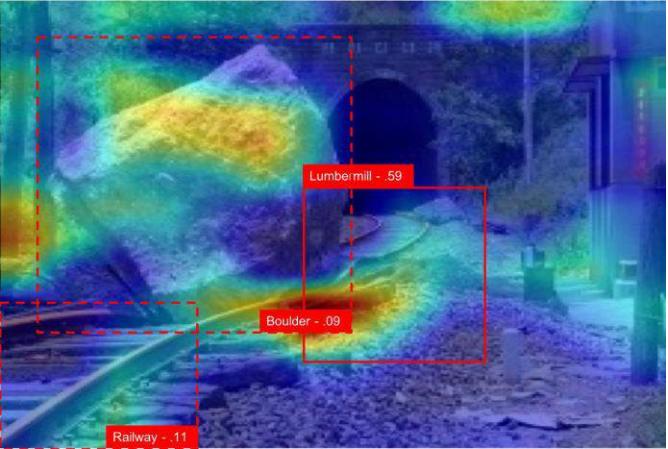
Boulder - .09

A screenshot of the DBpedia 'About: Boulder' page. The title 'About: Boulder' is at the top, followed by a brief description: '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 bulerston or Swedish bollersten. In places covered by ice sheets during Ice Age, such as Scandinavia, northern North America, and Russia, glacial erratics are common. Erratic 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 Hobbiton battlefields in New Zealand, where an entire valley contains only boulders, and The Boulders 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.' Below this is a table of properties and values:

Property	Value
db:abstract	<ul style="list-style-type: none">In geology, a boulder is a rock fragment with size greater than 25.6 centimetres (10.1 in) in diameter. Smaller pieces are called cobbles and pebbles, depending on their "grain size". While a boulder may be small enough to move or roll manually, others are extremely massive. In common usage, a boulder is too large for a person to move. Smaller boulders are usually just called rocks or stones. The word boulder is short for boulder stone, from Middle English bulerston or Swedish bollersten. In places covered by ice sheets during Ice Age, such as Scandinavia, northern North America, and Russia, glacial erratics are common. Erratic 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 Hobbiton battlefields in New Zealand, where an entire valley contains only boulders, and The Boulders 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.
dbo:thumbnail	<ul style="list-style-type: none">wiki-commons/Special:File?path=Boulderized_Rock.jpg&wheight=300
dbo:wikiPageID	<ul style="list-style-type: none">60784 (edit)
dbo:wikiPageRevisionID	<ul style="list-style-type: none">743049914 (edit)
db:subject	<ul style="list-style-type: none">db:Rock_formationsdb:Rocks

A screenshot of the DBpedia 'About: Rail transport' page. The title 'About: Rail transport' is at the top, followed by a brief description: '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.' Below this is a table of properties and values:

Property	Value
db:abstract	<ul style="list-style-type: none">Rail transport is a means of conveyance of passengers and goods on wheeled vehicles running on rails, also known as tracks. It is also commonly referred to as train transport. In contrast to road transport, where vehicles run on a prepared flat surface, rail vehicles (rolling stock) are directionally guided by the tracks on which they run. Tracks usually consist of steel rails, installed on ties (sleepers) and ballast, on which the rolling stock, usually fitted with metal wheels, moves. Other variations are also possible, such as slab track, where the rails are fastened to a concrete foundation resting on a prepared subsurface. Rolling stock in a rail transport system generally encounters lower frictional resistance than road vehicles, so passenger and freight cars (carriages and wagons) can be coupled into longer trains. The operation is carried out by a railway company, providing transport between train stations or freight customer facilities. Power is provided by locomotives which either draw electric power from a railway electrification system or produce their own power usually by diesel engines. Most trains 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 Persepolis, one of the Seven Wonders of Greece,



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

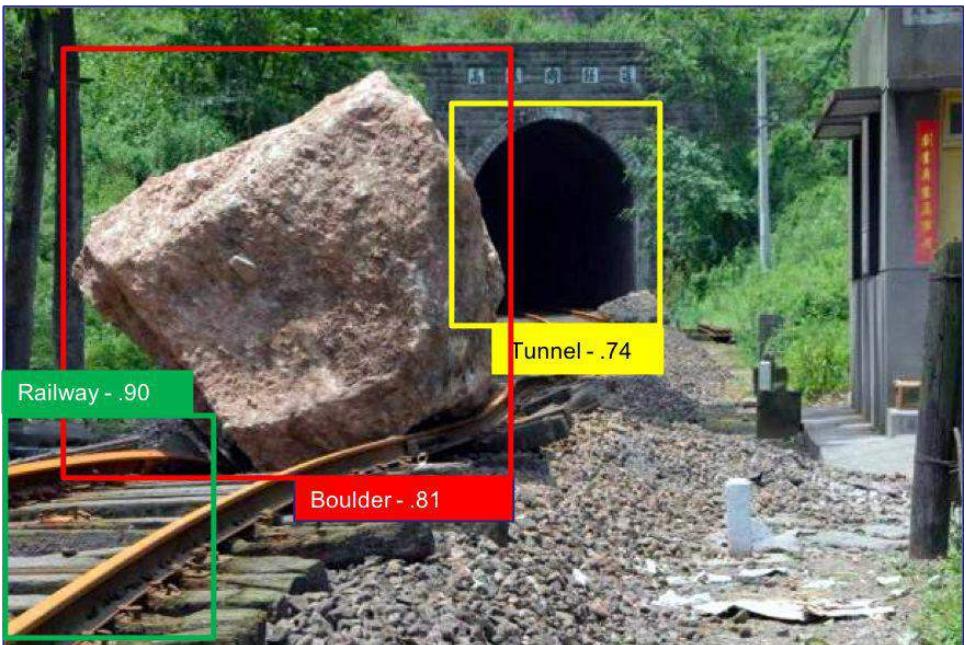
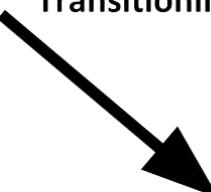
- **Hardware:** High performance, scalable, generic (to different FGPA family) & portable CNN dedicated programmable processor implemented on an FPGA for **real-time embedded inference**



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



Transitioning



EXPLANATIONS

ResNet50 image classifier



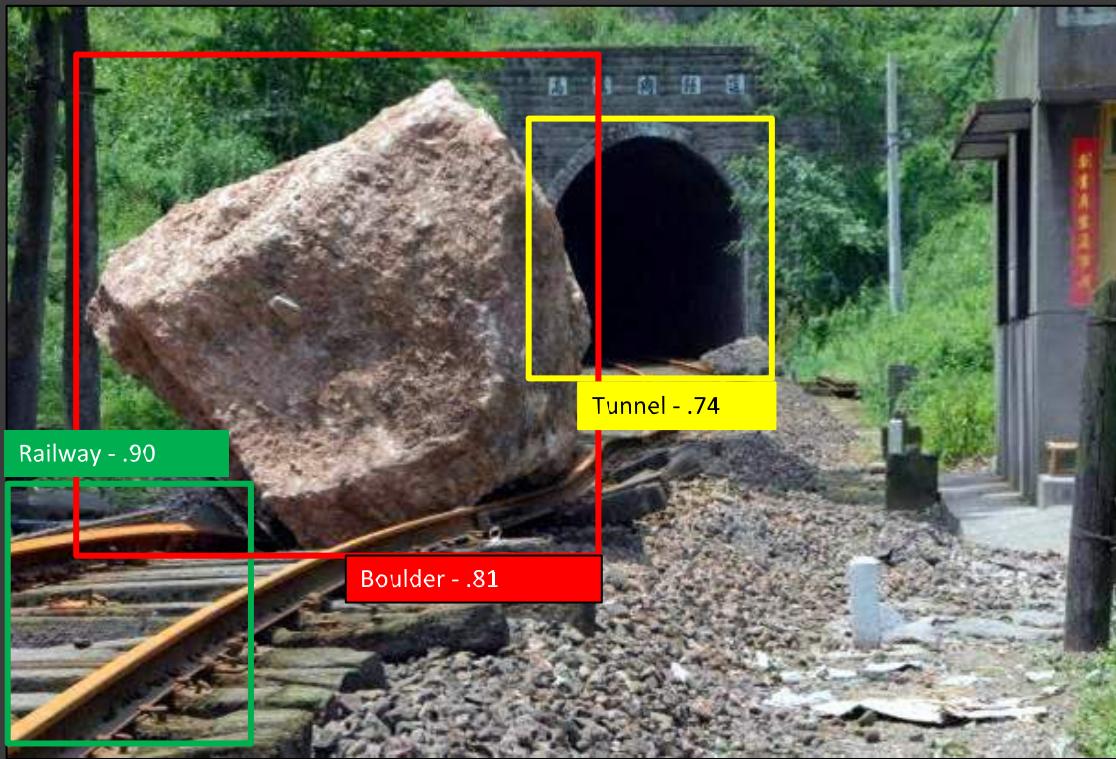
operating
on
obstructing

Rail
Trac
k
Boulder

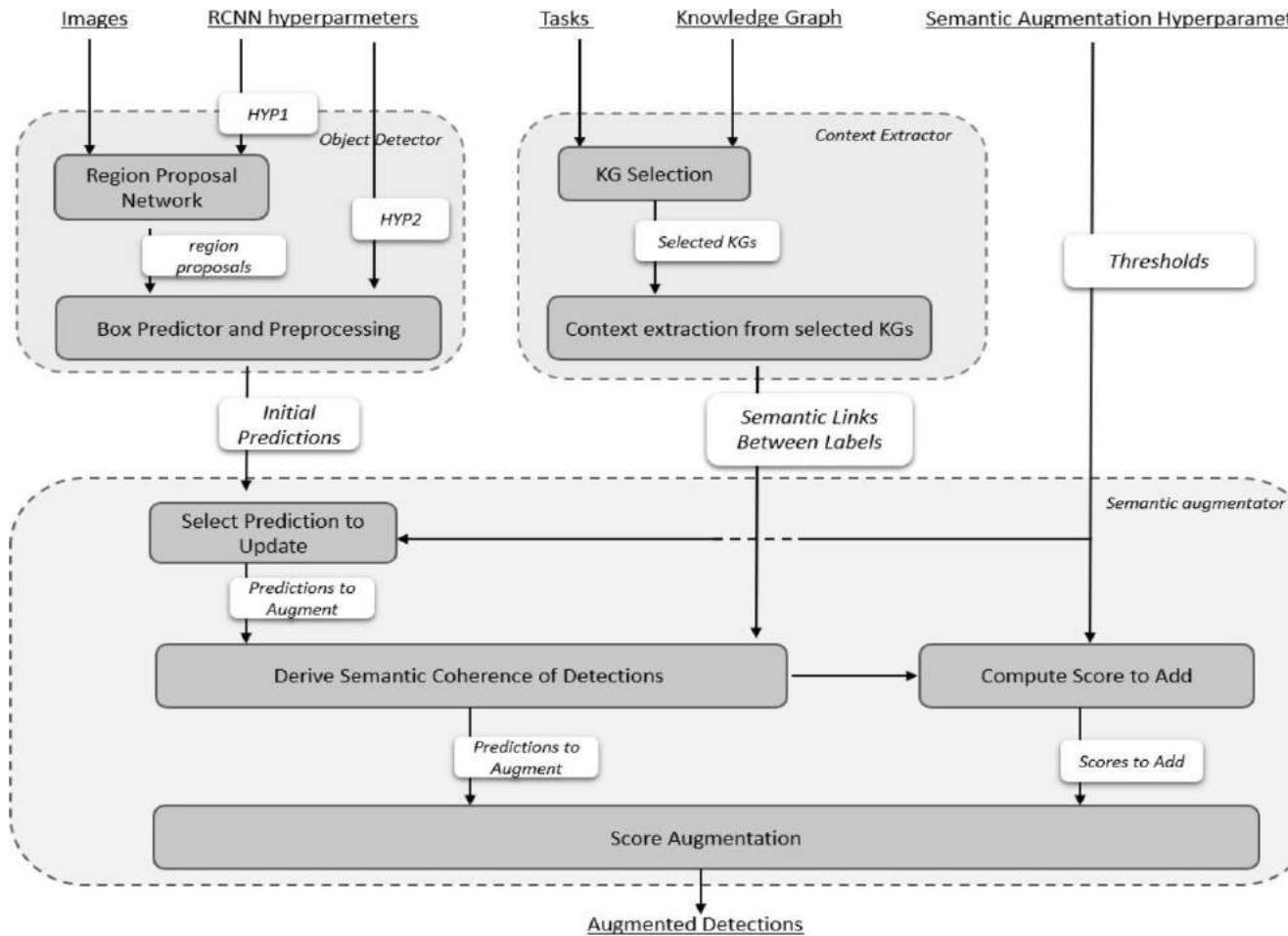
Landslide

Obstacle

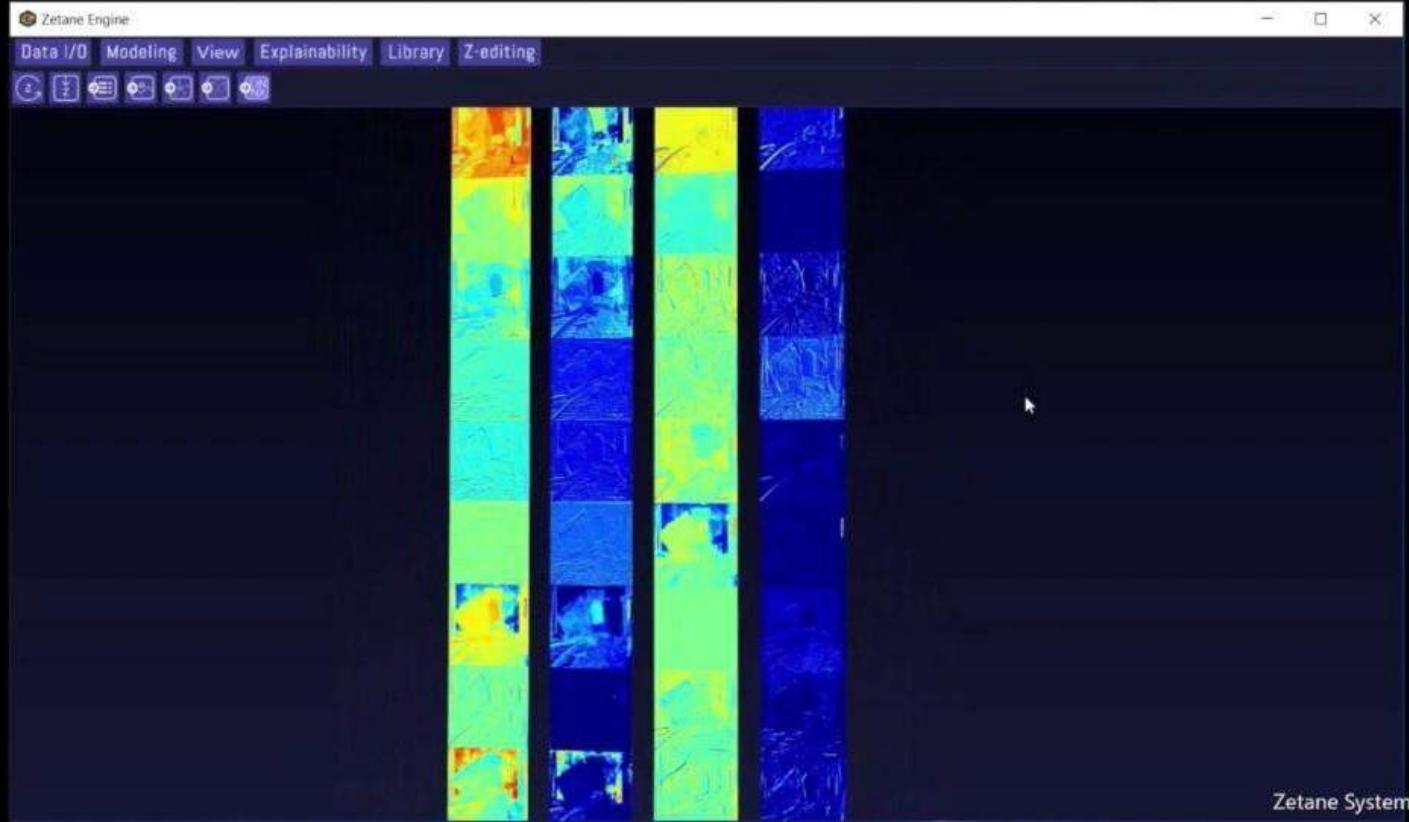
Tunne
l



Knowledge Graph in Machine Learning - An Implementation



XAI Tools on Applications, Lessons Learnt and Research Challenges



Explainable Boosted Object Detection – Industry Agnostic

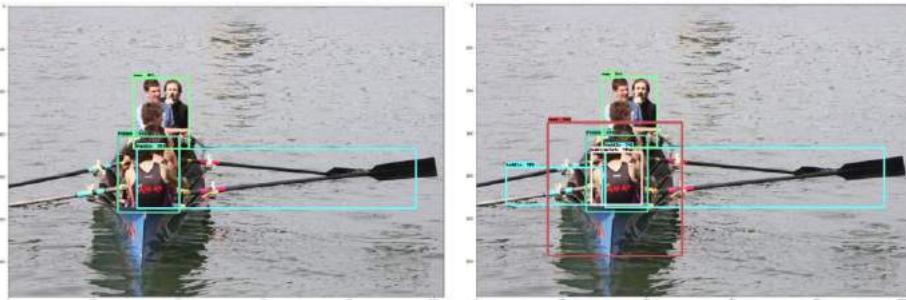
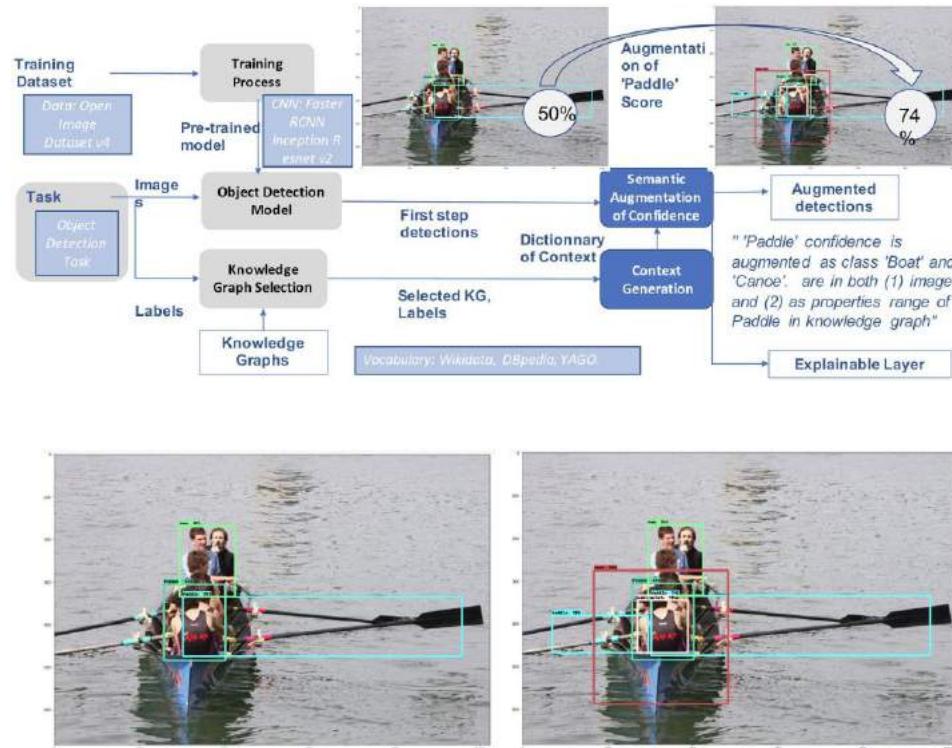


Fig. 2. Left image: results from baseline Faster RCNN: Paddle: 50% confidence, Person: 66%, Man: 46%. Right image: results from the semantic augmentation: **Paddle: 74% confidence, Person: 66%, Man: 56%, Boat: 58%** with explanation: Person, Paddle, Water as part of the context in the image and knowledge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

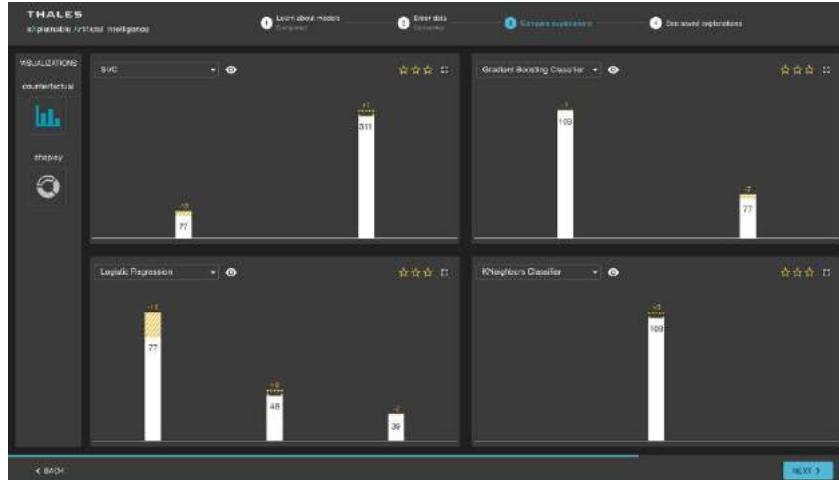
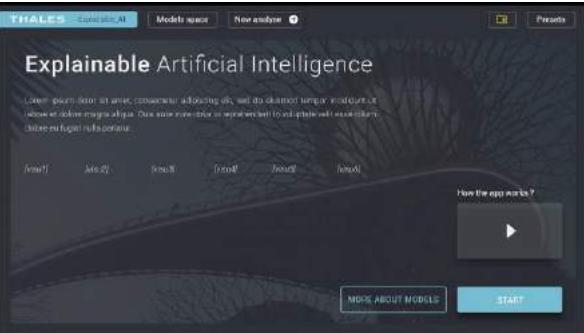
Goal

- 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



EXPLANATIONS

ResNet50 image classifier

Prediction: tank (id:4389033) with proba:
0.8574951887130737

Lime

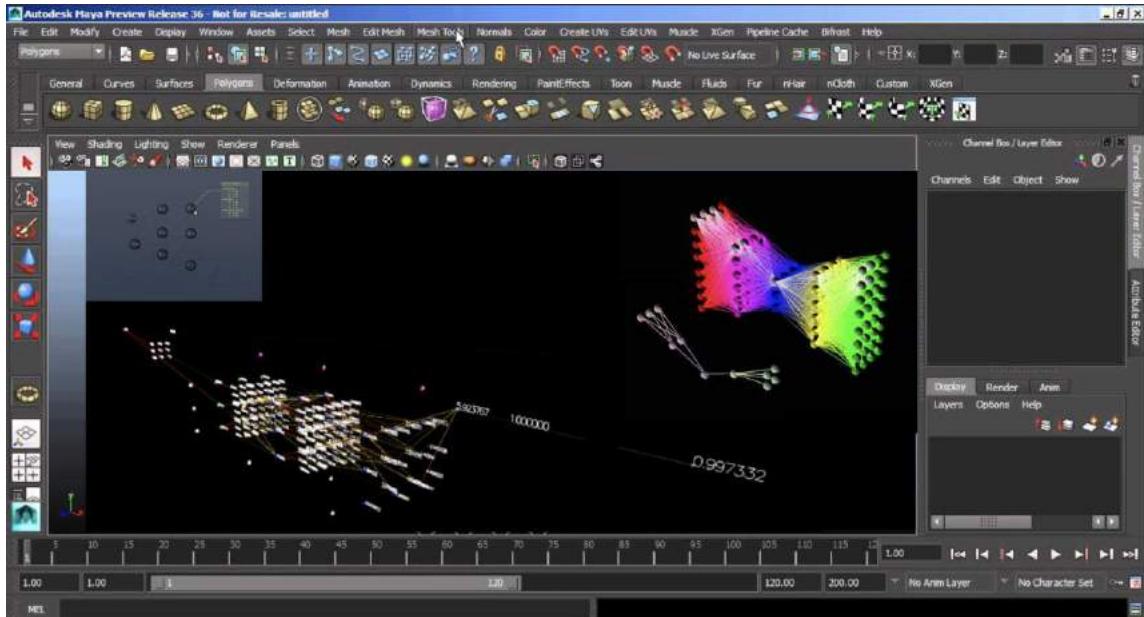


#1: Explaining Image Classification

Data: Image – XAI: Saliency Masks

PetalFocus: 1110000

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.

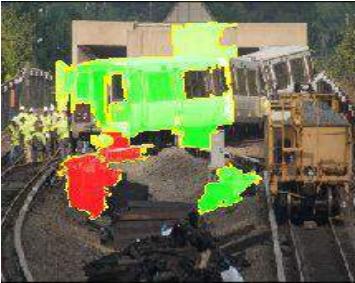
AI Technology: Artificial Neural Network

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

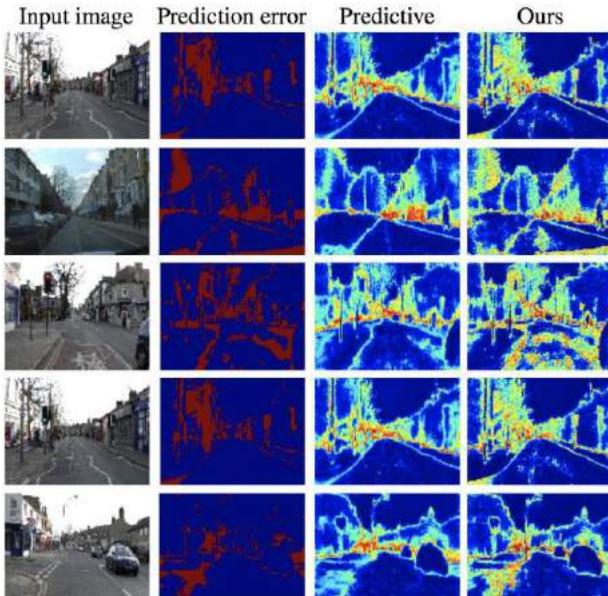


Zetane.com

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



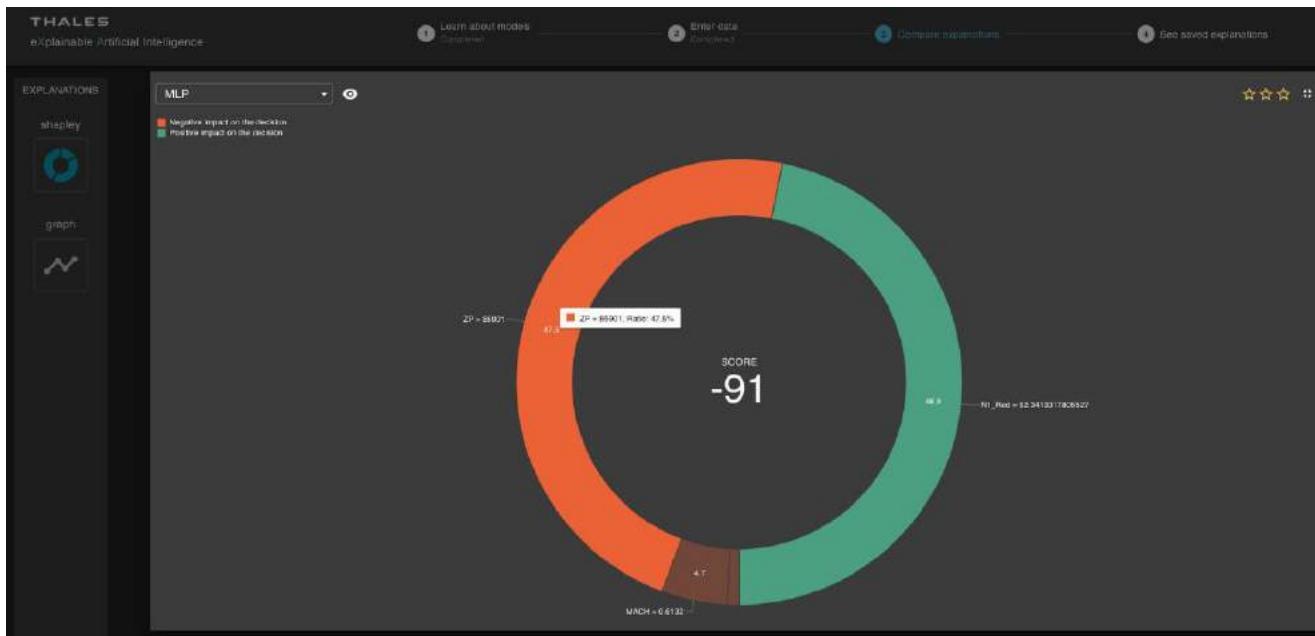
Explaining Flight Performance- Transportation

Challenge: Predicting and explaining aircraft engine performance

AI Technology: Artificial Neural Networks

XAI Technology: Shapely Values

THALES



Explainable On-Time Performance - Transportation

KLM / Transavia Flight Delay Prediction												
PLANE INFO	ARRIVAL				TURNAROUND			DEPARTURE				
	Status / Aircraft	Flight	ETA	Status	Delay Code	Gate	Slot	Progress	Milestones	Flight	ETA	Status
✓ <i>urtwet</i> ✓	4567	18:30	Scheduled	-	345345	1	<div style="width: 50%; background-color: green;"></div>	<div style="width: 50%; background-color: grey;"></div>	5678	19:00	Scheduled	-
⚠ <i>ldafew</i> ✓	4567	18:30	Delayed	ABC, DEF, GHI	345345	1	<div style="width: 10%; background-color: red;"></div>	<div style="width: 90%; background-color: grey;"></div>	5678	19:00	Delayed	ABC, DEF, GHI
✓ <i>psidib</i> ✓	4567	18:30	Scheduled	ABC, DEF, GHI	345345	1	<div style="width: 50%; background-color: green;"></div>	<div style="width: 50%; background-color: grey;"></div>	5678	19:00	Scheduled	ABC, DEF, GHI
🚫 <i>kahdbs</i> ✓	4567	-	Cancelled	ABC, DEF, GHI	-	-	<div style="width: 0%; background-color: grey;"></div>	<div style="width: 100%; background-color: grey;"></div>	5678	-	Cancelled	ABC, DEF, GHI
⚠ <i>scoradis</i> ✘	4567	18:35	Delayed	ABC, DEF, GHI	345345	1	<div style="width: 20%; background-color: yellow;"></div>	<div style="width: 80%; background-color: grey;"></div>	5678	19:00	Delayed	ABC, DEF, GHI
⚠ <i>adobis</i> ✘	4567	18:30	Delayed	ABC, DEF, GHI	345345	1	<div style="width: 10%; background-color: red;"></div>	<div style="width: 90%; background-color: grey;"></div>	5678	19:00	Scheduled	ABC, DEF, GHI
✓ <i>aedbac</i> ✓	4567	18:30	Scheduled	ABC, DEF, GHI	345345	1	<div style="width: 50%; background-color: green;"></div>	<div style="width: 50%; background-color: grey;"></div>	5678	19:00	Scheduled	ABC, DEF, GHI
✓ <i>aedbac</i> ✓	4567	18:30	Scheduled	ABC, DEF, GHI	345345	1	<div style="width: 50%; background-color: green;"></div>	<div style="width: 50%; background-color: grey;"></div>	5678	19:00	Scheduled	ABC, DEF, GHI
✓ <i>aedbac</i> ✓	4567	18:30	Scheduled	ABC, DEF, GHI	345345	1	<div style="width: 50%; background-color: green;"></div>	<div style="width: 50%; background-color: grey;"></div>	5678	19:00	Scheduled	ABC, DEF, GHI
✓ <i>aedbac</i> ✓	4567	18:30	Scheduled	ABC, DEF, GHI	345345	1	<div style="width: 50%; background-color: green;"></div>	<div style="width: 50%; background-color: grey;"></div>	5678	19:00	Scheduled	ABC, DEF, GHI
✓ <i>aedbac</i> ✓	4567	18:30	Scheduled	ABC, DEF, GHI	345345	1	<div style="width: 50%; background-color: green;"></div>	<div style="width: 50%; background-color: grey;"></div>	5678	19:00	Scheduled	ABC, DEF, GHI
✓ <i>aedbac</i> ✓	4567	18:30	Scheduled	ABC, DEF, GHI	345345	1	<div style="width: 50%; background-color: green;"></div>	<div style="width: 50%; background-color: grey;"></div>	5678	19:00	Scheduled	ABC, DEF, GHI
✓ <i>aedbac</i> ✓	4567	18:30	Scheduled	ABC, DEF, GHI	345345	1	<div style="width: 50%; background-color: green;"></div>	<div style="width: 50%; background-color: grey;"></div>	5678	19:00	Scheduled	ABC, DEF, GHI
✓ <i>aedbac</i> ✓	4567	18:30	Scheduled	ABC, DEF, GHI	345345	1	<div style="width: 50%; background-color: green;"></div>	<div style="width: 50%; background-color: grey;"></div>	5678	19:00	Scheduled	ABC, DEF, GHI
✓ <i>aedbac</i> ✓	4567	18:30	Scheduled	ABC, DEF, GHI	345345	1	<div style="width: 50%; background-color: green;"></div>	<div style="width: 50%; background-color: grey;"></div>	5678	19:00	Scheduled	ABC, DEF, GHI
✓ <i>aedbac</i> ✓	4567	18:30	Scheduled	ABC, DEF, GHI	345345	1	<div style="width: 50%; background-color: green;"></div>	<div style="width: 50%; background-color: grey;"></div>	5678	19:00	Scheduled	ABC, DEF, GHI
✓ <i>aedbac</i> ✓	4567	18:30	Scheduled	ABC, DEF, GHI	345345	1	<div style="width: 50%; background-color: green;"></div>	<div style="width: 50%; background-color: grey;"></div>	5678	19:00	Scheduled	ABC, DEF, GHI
✓ <i>aedbac</i> ✓	4567	18:30	Scheduled	ABC, DEF, GHI	345345	1	<div style="width: 50%; background-color: green;"></div>	<div style="width: 50%; background-color: grey;"></div>	5678	19:00	Scheduled	ABC, DEF, GHI
✓ <i>aedbac</i> ✓	4567	18:30	Scheduled	ABC, DEF, GHI	345345	1	<div style="width: 50%; background-color: green;"></div>	<div style="width: 50%; background-color: grey;"></div>	5678	19:00	Scheduled	ABC, DEF, GHI
✓ <i>aedbac</i> ✓	4567	18:30	Scheduled	ABC, DEF, GHI	345345	1	<div style="width: 50%; background-color: green;"></div>	<div style="width: 50%; background-color: grey;"></div>	5678	19:00	Scheduled	ABC, DEF, GHI

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.

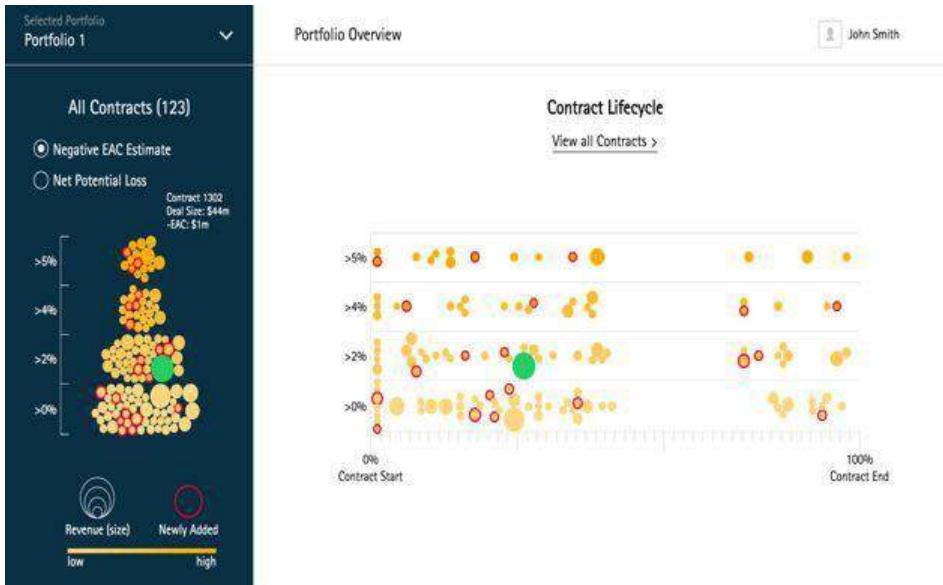
XAI Technology: Knowledge graph embedded Sequence Learning using LSTMs



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

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

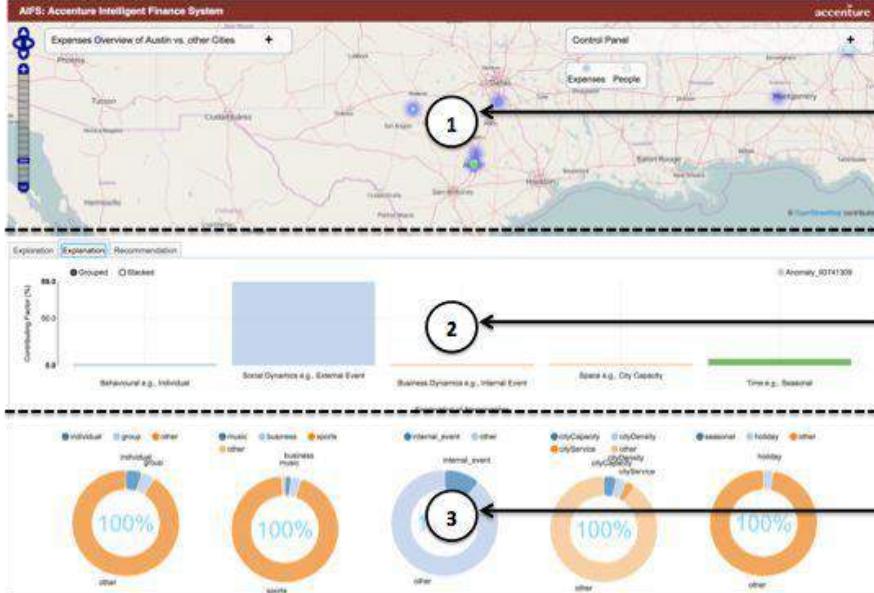


Challenge: Accenture is managing every year more than 80,000 opportunities and 35,000 contracts with an expected revenue of \$34.1 billion. Revenue expectation does not meet estimation due to the complexity and risks of critical contracts. This is, in part, due to the (1) large volume of projects to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

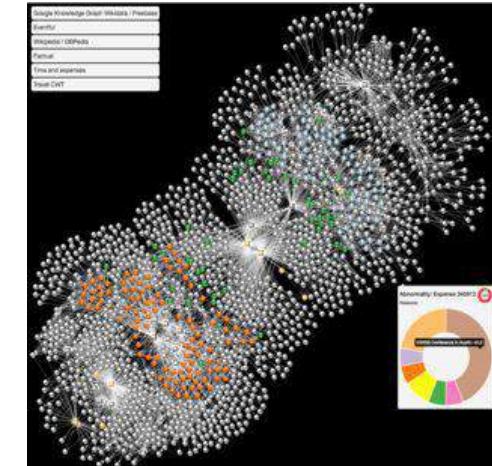
Explainable Anomaly Detection – Finance (Compliance)



Data analysis
for spatial interpretation
of abnormalities:
abnormal expenses

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

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



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

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

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

XAI Technology: Knowledge graph embedded Ensemble Learning

Counterfactual Explanations for Credit Decisions (3) - Finance



Sorry, your loan application has been rejected.

Our analysis:

The following features were too high:

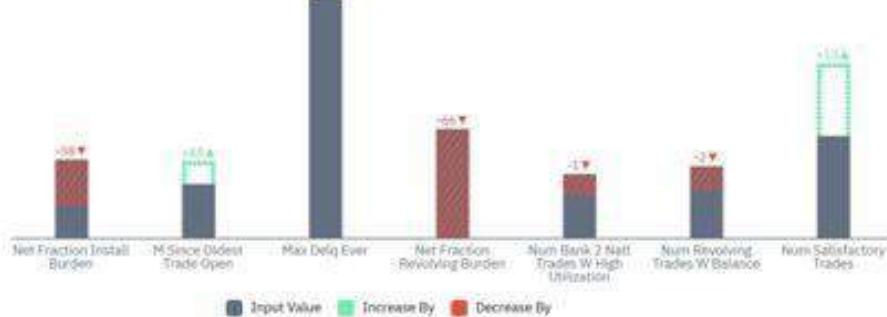
PercentInstallTrad... NetFractionRevolv... NetFractionInstall...
NumRevolvingTra... NumBank2NatTra... PercentTradesWB...

The following features were too low:

MSinceOldestTrad... AverageMinFile... NumTotalTrades...

The following features require changes:

MaxDelq2PublicC... MaxDelqEver



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

Explanation of Medical Condition Relapse – Health



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

LinkedIn™ Talent Search Case Study: **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

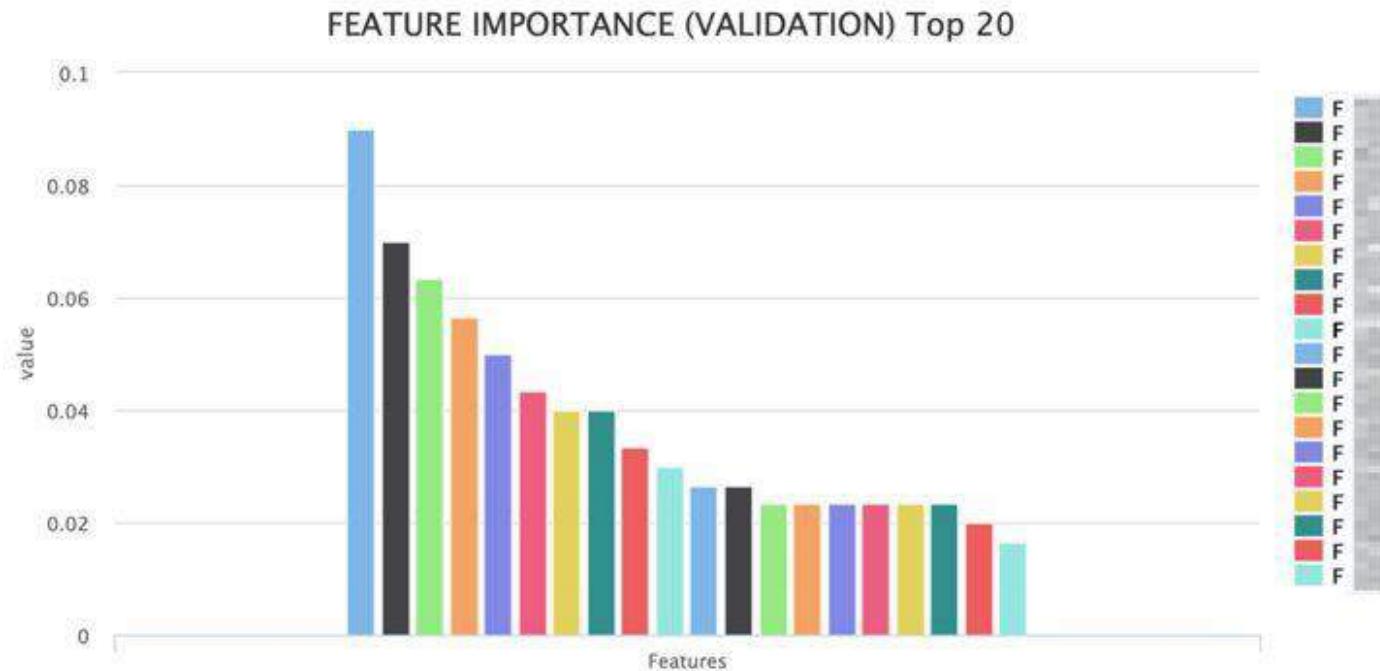
The screenshot shows the LinkedIn Recruiter software interface. At the top, there's a navigation bar with 'RECRUITER' and links for 'PROJECTS', 'CLIPBOARD', 'JOBS', and 'REPORTS'. Below the navigation is a search bar with a magnifying glass icon. To the right of the search bar are three numerical statistics: '1,767,429 total candidates', '216,022 are more likely to respond', and '161,334 open to new opportunities'. On the left, there's a sidebar with filtering options: 'SHOWING DATA FOR', 'Title' (with 'User Experience Designer' selected), 'Skill' (with 'Interaction Designer' selected), 'Location' (with 'United States' selected), 'Industry' (selected), and 'Employment type' (selected). The main area displays a list of six candidate profiles, each with a small profile picture, name, title, current employer, location, and employment status. Each profile also has a 'More' link.

Name	Title	Employer	Location	Employment Status	More
Eloora Tyler	User Experience Designer	Flexis	Minneapolis, Minnesota • Accounting	2017 - Present	More
Carl Meyer	Product Designer	Flexis	Minneapolis, Minnesota • Accounting	2016 - Present	More
Alma Frazier	Interaction Designer	Eastern Fellows	Minneapolis, Minnesota • Accounting	2014 - Present	More
Ray Patterson	UX Designer	MJ Accountants	Minneapolis, Minnesota • Accounting	2013 - Present	More
Susie Jensen	UX Designer	Eastern Fellows	Minneapolis, Minnesota • Accounting	2014 - Present	More

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

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

- 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

Model Interpretation for Predictive Models in B2B Sales Predictions

Jilei Yang, Wei Di, Songtao Guo



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?

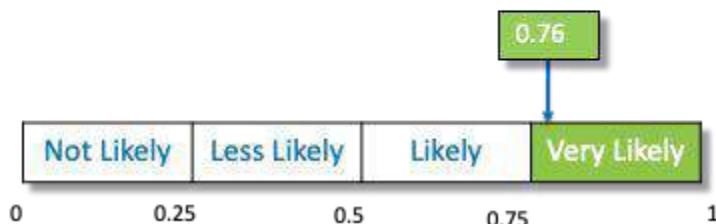
① Feature Contributor

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)



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

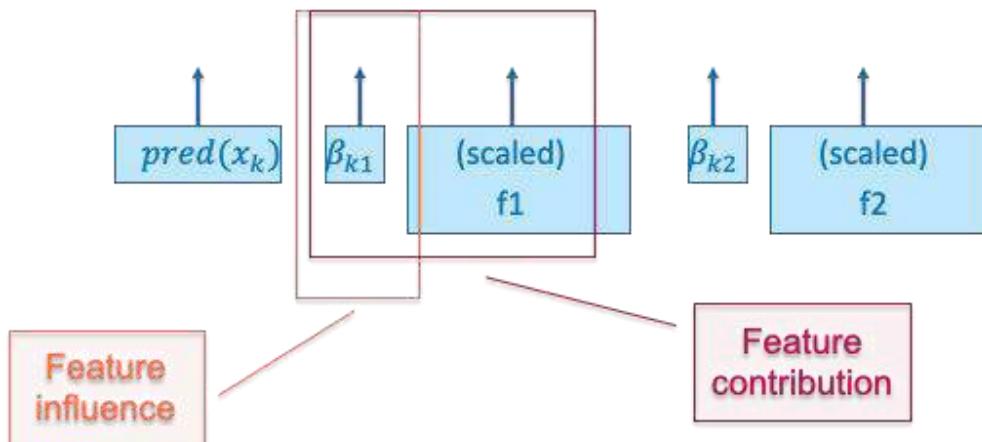
Revisiting LIME

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

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

- E.g., for company CompanyX:

$$0.76 \approx 1.82 * 0.17 + 1.61 * 0.11 + \dots$$



xLIME

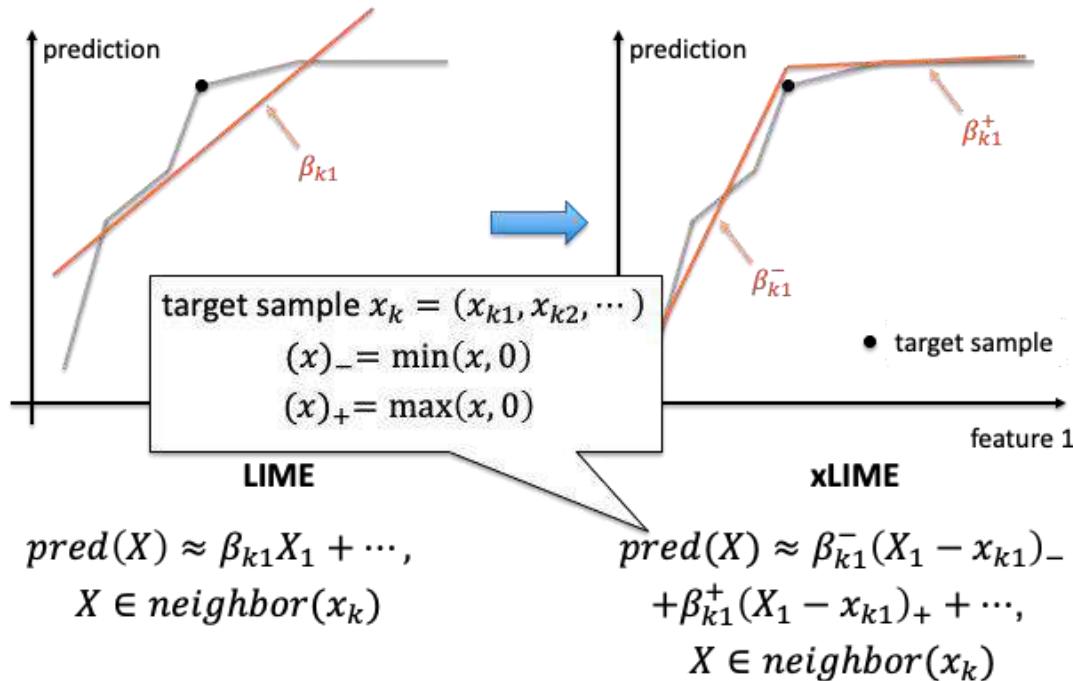
Piecewise Linear
Regression

Localized Stratified
Sampling



Piecewise Linear Regression

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

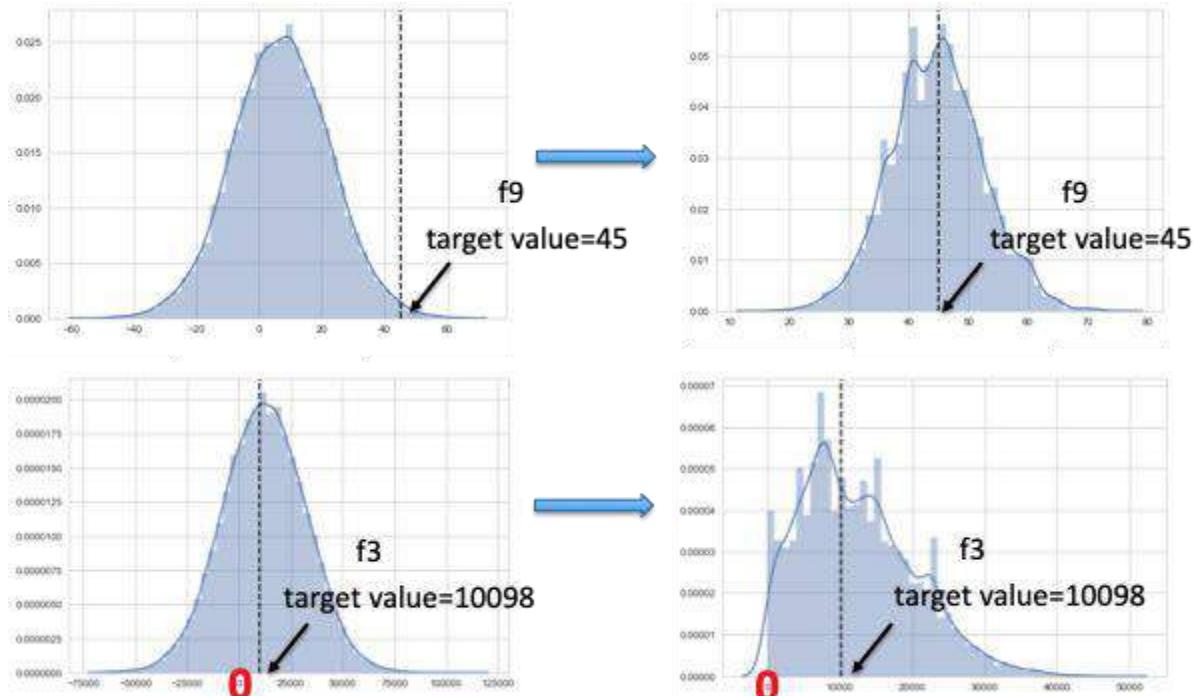


Impact of Piecewise Approach

- Target sample $x_k = (x_{k1}, x_{k2}, \dots)$
- 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



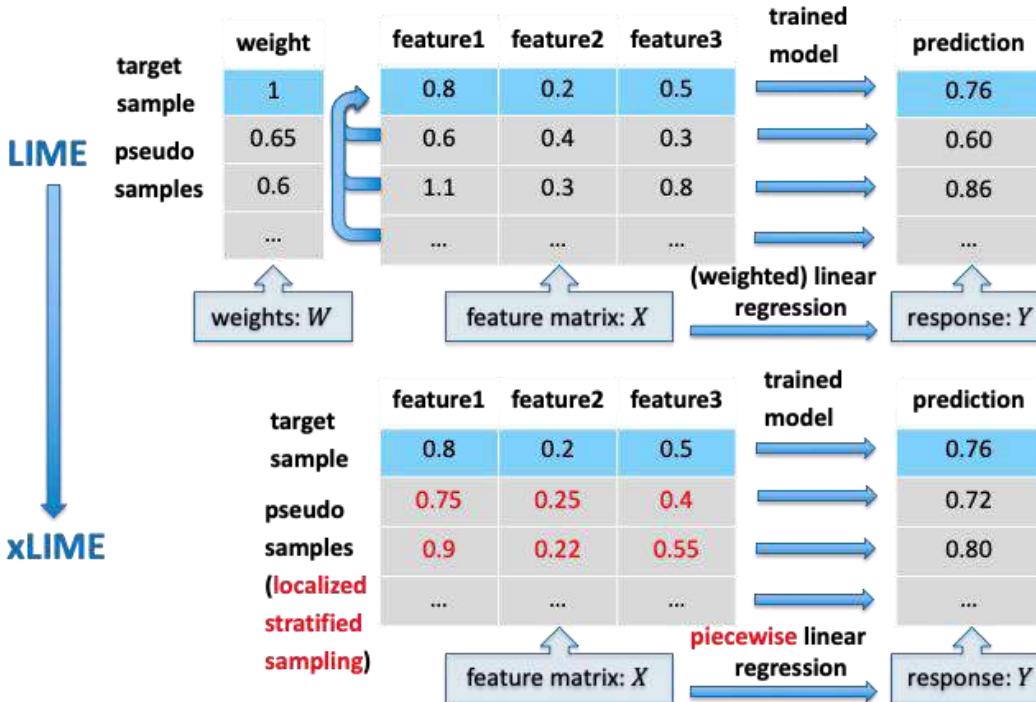
Localized Stratified Sampling: Method

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

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

- $p_j(X_j)$: 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 $N(x_j, s_j^2)$.

Summary



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

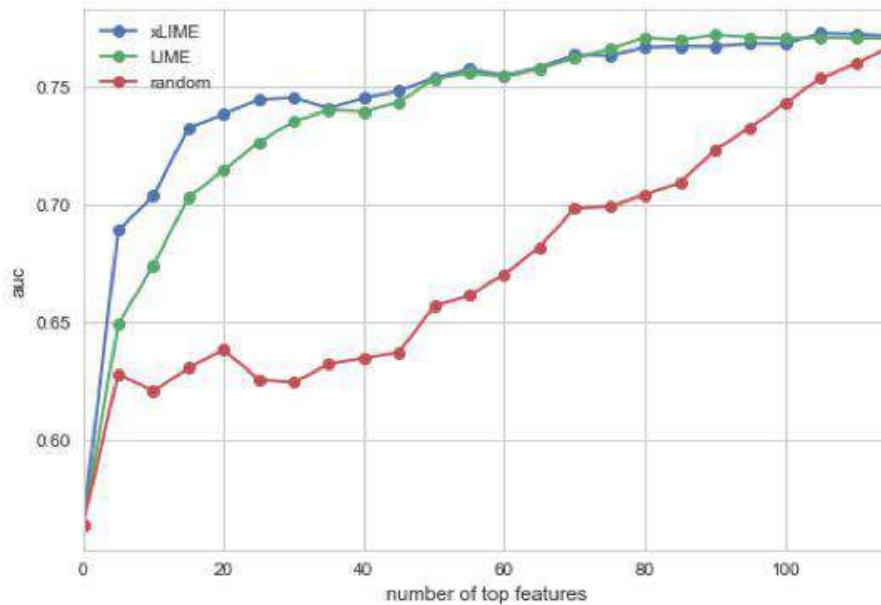
LIME

	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

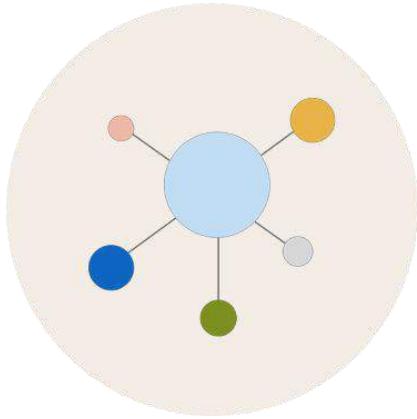
Company: CompanyX

	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

Debugging Relevance Models



Modeling

Improve the machine learning model



Value

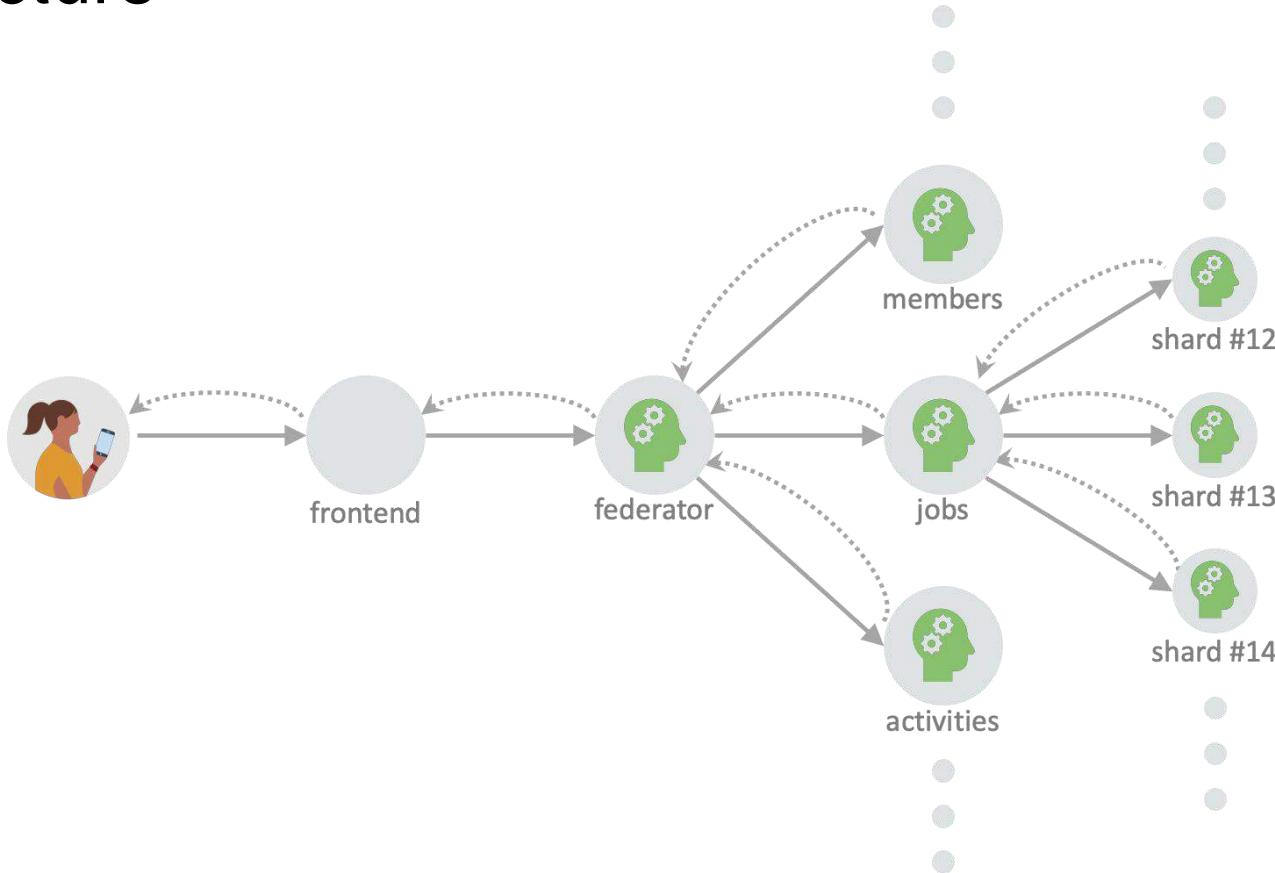
Bring value to our members
by providing relevant
experience



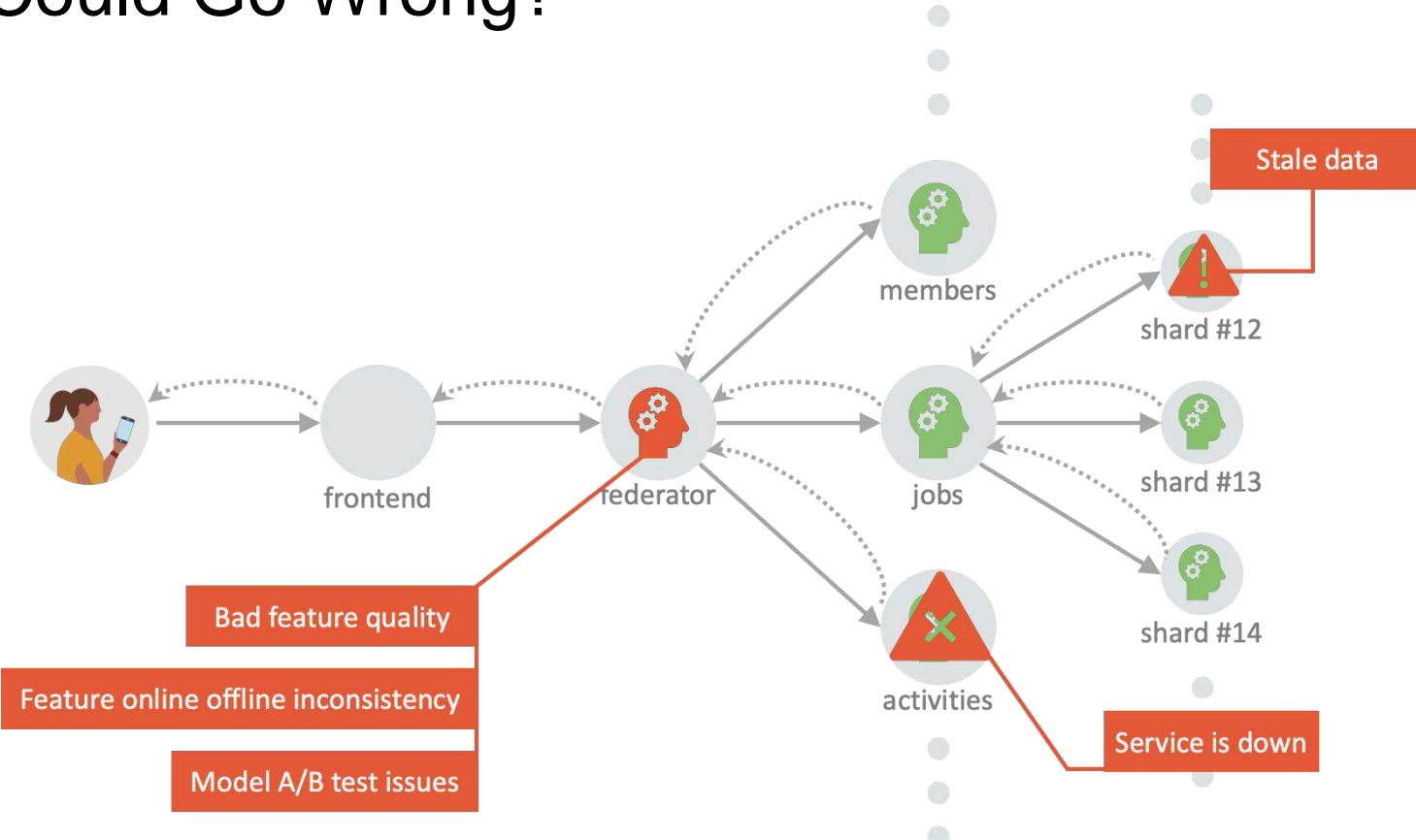
Trust

Build trust with our members

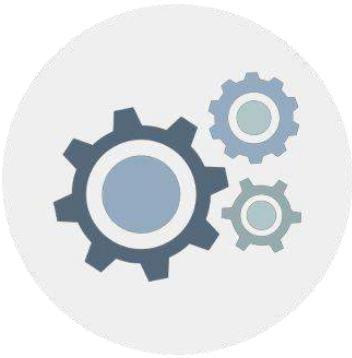
Architecture



What Could Go Wrong?



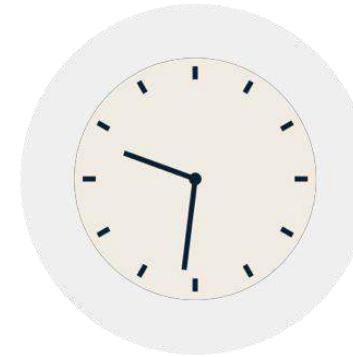
Challenges



Complex Infrastructure



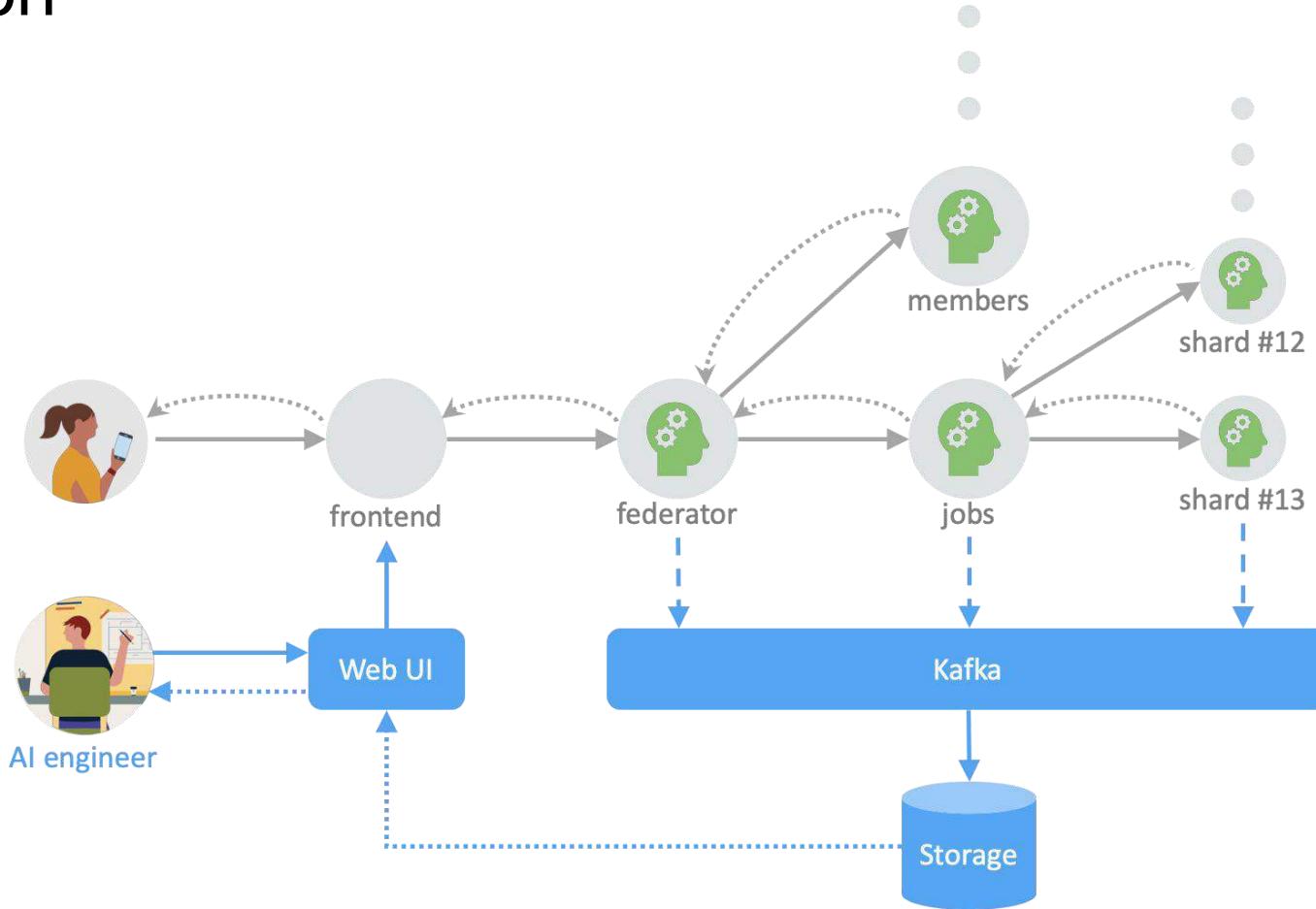
Hard to Reproduce



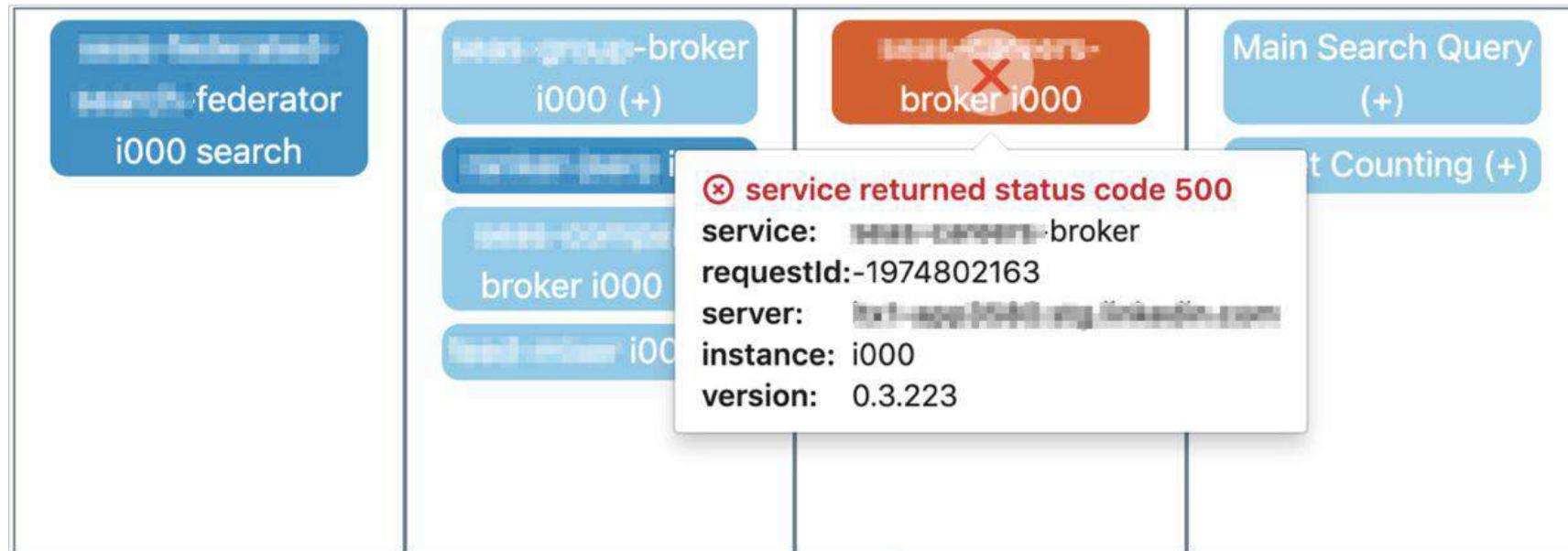
Time Consuming



Solution



Call Graph



Results

Request

Response

Host Information

Why Not Seen

Logs

① FPR task(s) failed: 1

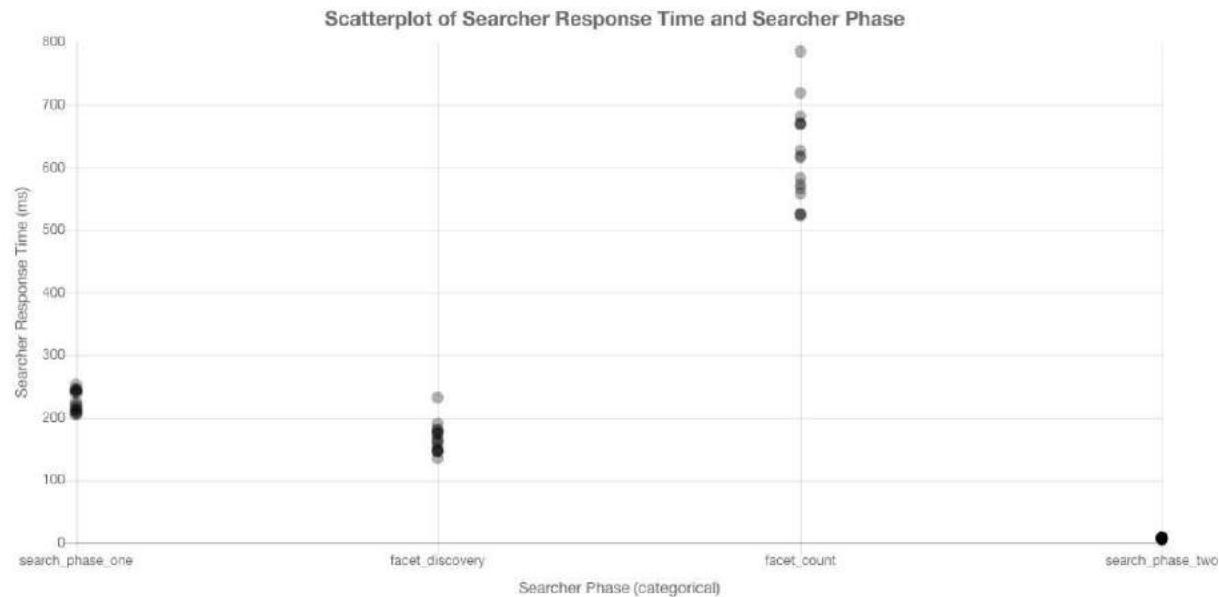
① Cannot adapt response from fpr, adapter: [REDACTED], Service: [REDACTED], ResourceMethod: FINDER, Cause: task: [REDACTED] withTimeout 1000ms

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



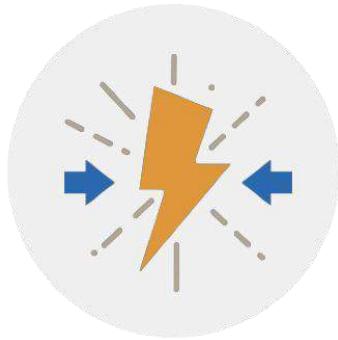
Features

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

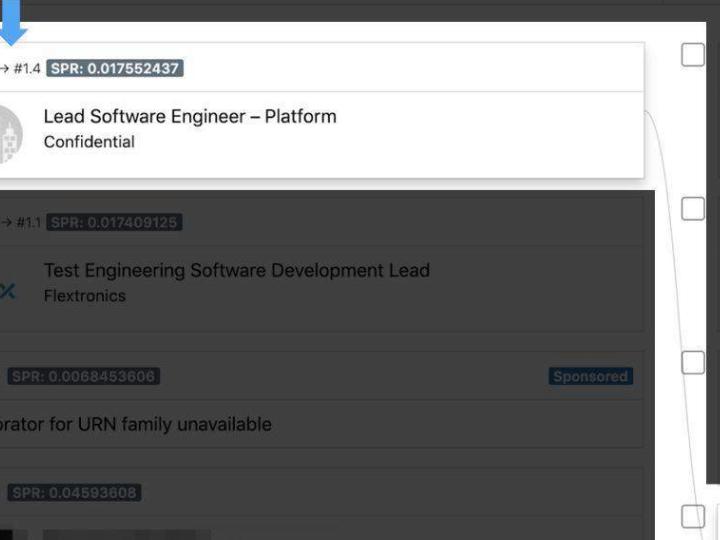
Compare features and scores of 2 different items, from the same query or different queries

Holistic Comparison

Position changes: 3 | New items: 11

Click to view details, or select to compare.

Query 1	cURL	Calltree	Query 2	cURL	Calltree
<p>#1.1 → #1.4 SPR: 0.017552437</p> <p> Lead Software Engineer – Platform Confidential</p>			<p>#1.2 → #1.1 SPR: 7.223955E-4</p> <p> Test Engineering Software Development Lead Flextronics</p>		
<p>#1.2 → #1.1 SPR: 0.017409125</p> <p> Test Engineering Software Development Lead Flextronics</p>			<p>→ #1.2 SPR: 6.688792E-4</p> <p> Software Engineer - Application Backend Yelp</p>		
<p>#2 → SPR: 0.0068453606</p> <p>Sponsored</p> <p>Decorator for URN family unavailable</p>			<p>→ #1.3 SPR: 6.687663E-4</p> <p> Software Engineer - Messaging Services Twilio</p>		
<p>#3 → SPR: 0.04593608</p> <p></p>			<p>#1.1 → #1.4 SPR: 6.686083E-4</p> <p> Lead Software Engineer – Platform Confidential</p>		
<p>#4 → SPR: 0.02149224</p>					



A screenshot of a search results comparison interface. At the top, it shows 'Position changes: 3 | New items: 11' and 'Click to view details, or select to compare.' Below this, there are two main sections: 'Query 1' on the left and 'Query 2' on the right. Each section contains a list of search results. A blue arrow points from the top result in Query 1 down to the top result in Query 2, indicating a direct comparison between these specific items. The results are presented in a table format with columns for each query and their respective cURL and Calltree links. The results themselves are listed in rows, showing job titles, companies, and scores (SPR). Some results are marked as 'Sponsored'. The overall layout is clean and organized, designed for easy comparison between two different search queries.

Granular Comparison

Query 1

Test Engineering Software Development Lead
Flextronics

Position	#1.2
Reference	urn:li:jobPosting:_____
SPR Score	0.017409125
Relevance Model	_____
Source Type	ORGANIC
FPR Model	_____

Query 2

Test Engineering Software Development Lead
Flextronics

Position	#1.1
Reference	urn:li:jobPosting:_____
SPR Score	7.2239555E-4
Relevance Model	_____
Source Type	ORGANIC
FPR Model	_____

All Groups

Search feature Shared features only Different values only

Group	Feature	Item 1	Item 2	% Change
SPR	responsePenalty /	4.0601455e-7	0.009018197	2221051.19
SPR	response	5.2125584e-9	0.000011580406	222063.57
SPR	score_response_viral	5.2125584e-9	0.000011580406	222063.57
SPR	diffHoursSinceLvFiveAndAgeInHour /	-3.0348454	-50.475624	1563.2

Replay

Feed Replay

Viewer ID
Viewer ID must be a LinkedIn employee.

Start Time (Pacific Time)
End Time (Pacific Time)

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:
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.)
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:
linkedin:like
urn:li:activity.
Relevance Model: nus:homepage_federator_relevance_463_ramp
FPR Model: m124_v2_multi_pass

4 urn:li:activity:
linkedin:react
urn:li:groupPost
Relevance Model: nus:homepage_federator_relevance_463_ramp

Teams

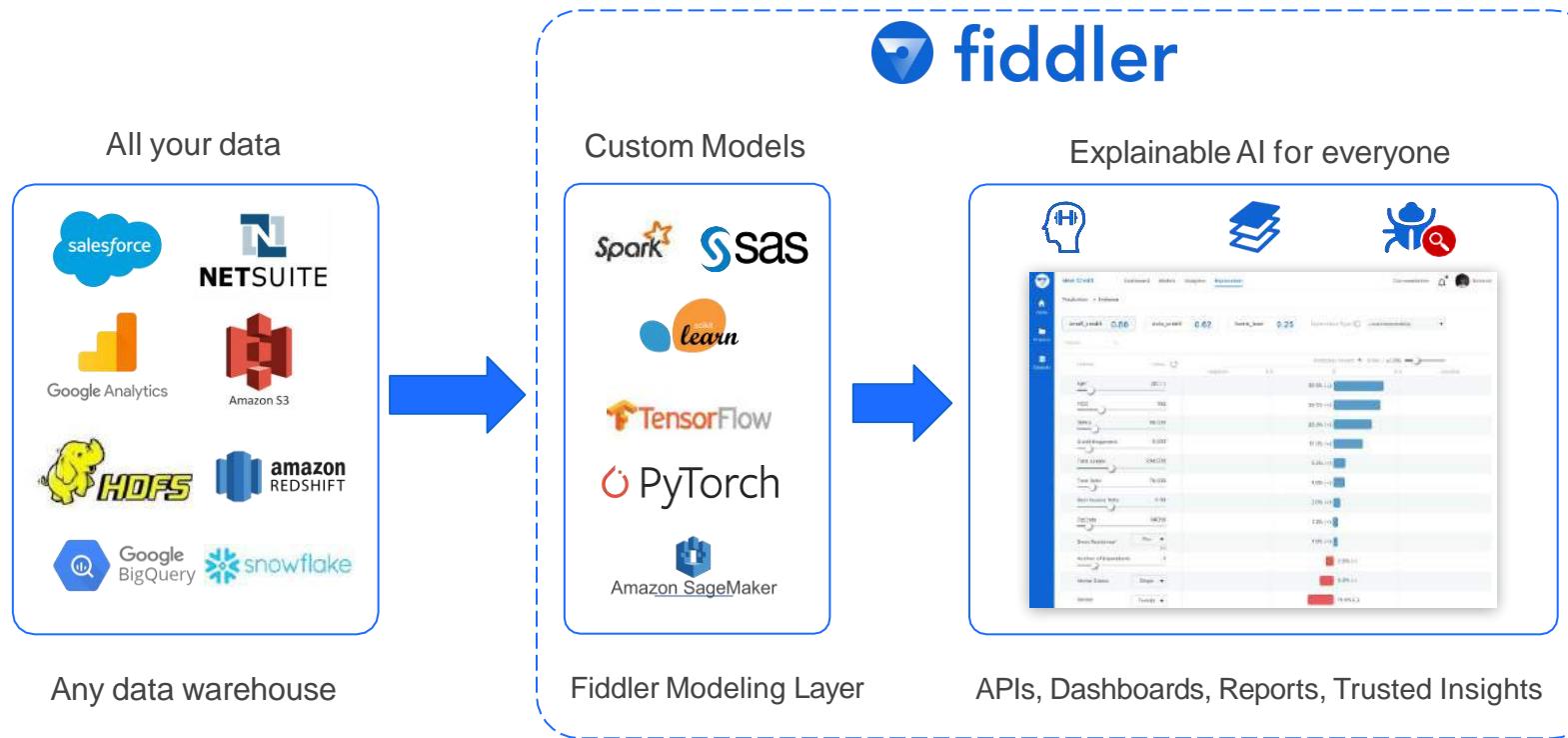
- Search
- Feed
- Comments
- People you may know
- Jobs you may be interested in
- Notification

Building an Explainable AI Engine @ fiddler

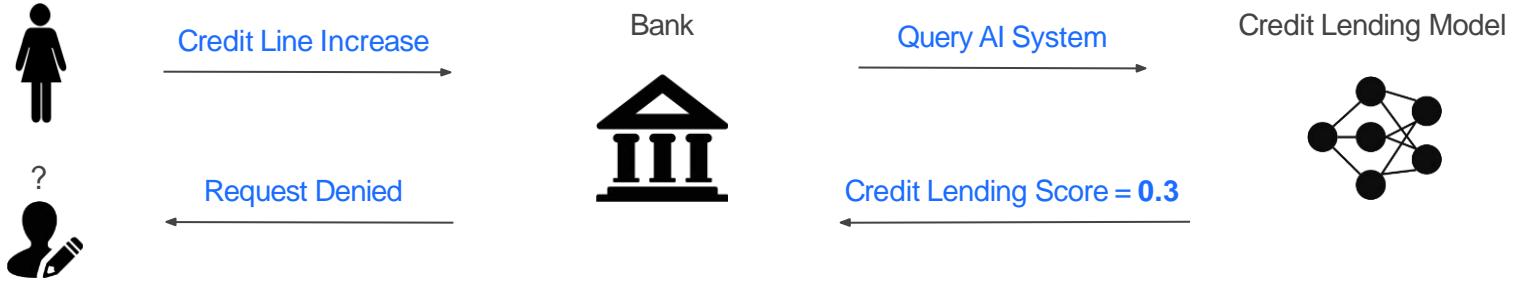
Case Study: Luke Merrick

Fiddler's Explainable AI Engine

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



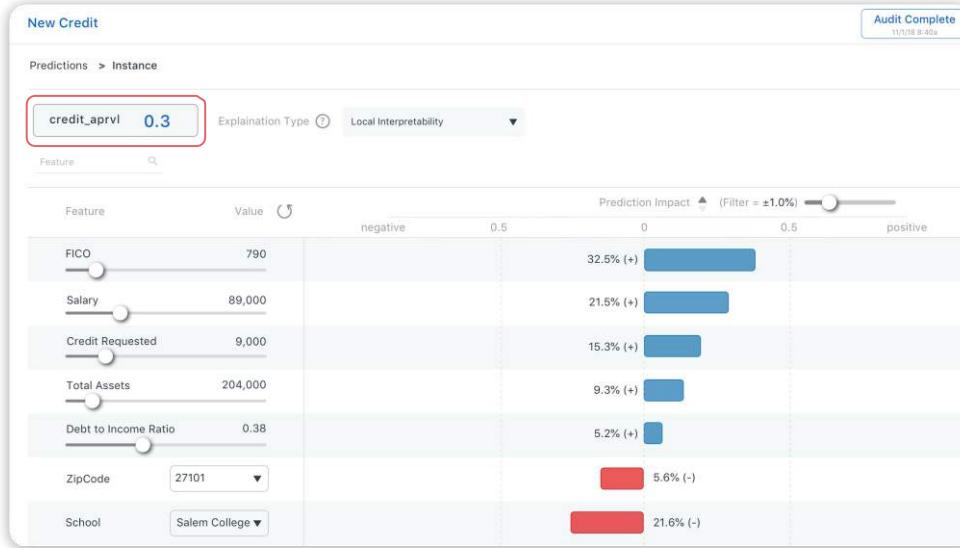
Example: Credit Lending in a black-box ML world



Why? Why not? How?

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

Explain individual predictions (using Shapley Values)



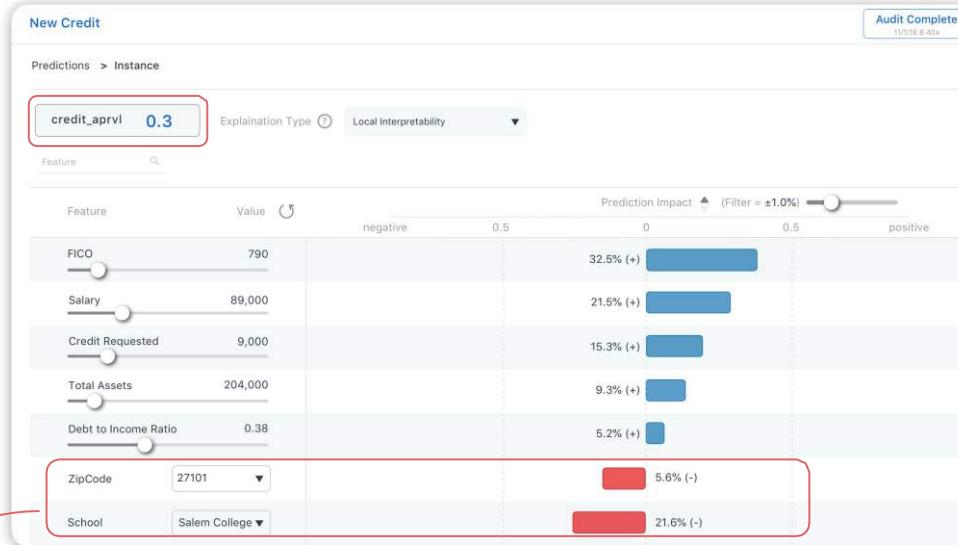
How Can This Help...

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)



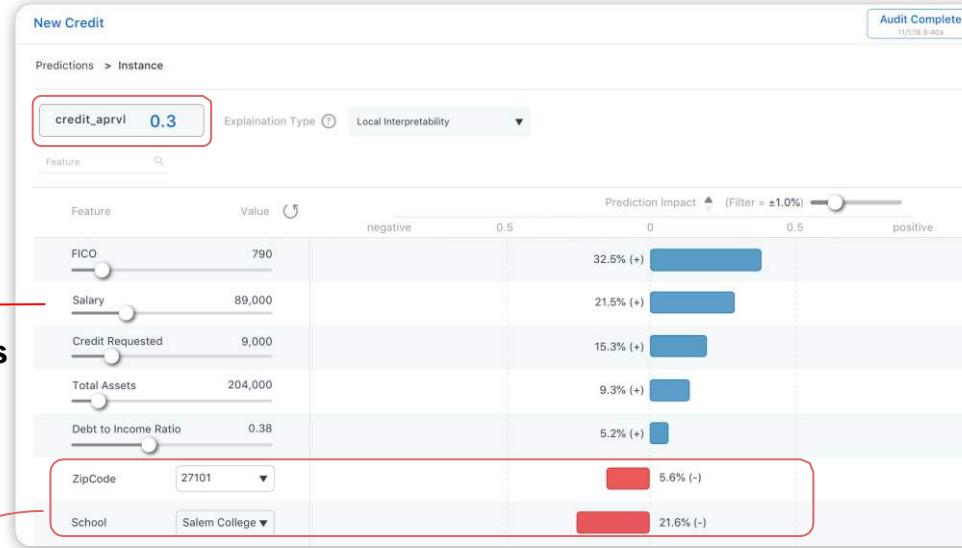
How Can This Help...

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

The screenshot shows the Alteryx Insights interface. On the left, a SQL query window displays a query that includes a slice operation:

```
1 /*
2 * EXAMPLES:
3 * example dataset query:
4 * select * from "your_dataset_name" limit 100
5 *
6 * example model query:
7 * select * from "your_dataset_name.your_model_name" limit 100
8 */
9
10 SLICE * from "p2p_loans.lingreg-all"
11 where "loan_amnt" < 10000
```

A red circle highlights the slice command. Below the query is a "Ready" button.

In the center, a "DATA" preview window shows a table with 13 rows of loan data. The first row is highlighted with a red box. A red arrow points from this row to the "EXPLANATION" card on the right.

The "EXPLANATION" card has a header "EXPLANATION" and "Impact ID: 37742142". It includes tabs for "Feature Impact" (which is selected), "Feature Correlation", and "Feature Distribution". It also has a "Top N Inputs" slider set to 10. The "Feature Impact" section lists features and their impact percentages:

Feature	Impact (%)
int_rate	14% (+)
dti	7% (+)
annual_inc	5% (+)
addr_state	3% (+)
iloc.range_low	3% (+)

A red arrow points from the "Feature Impact" card down to a smaller "Impact" card at the bottom right, which also shows the top 10 inputs:

Feature	Impact (%)
int_rate	14% (+)
dti	9% (+)

How Can This Help...

Global Explanations

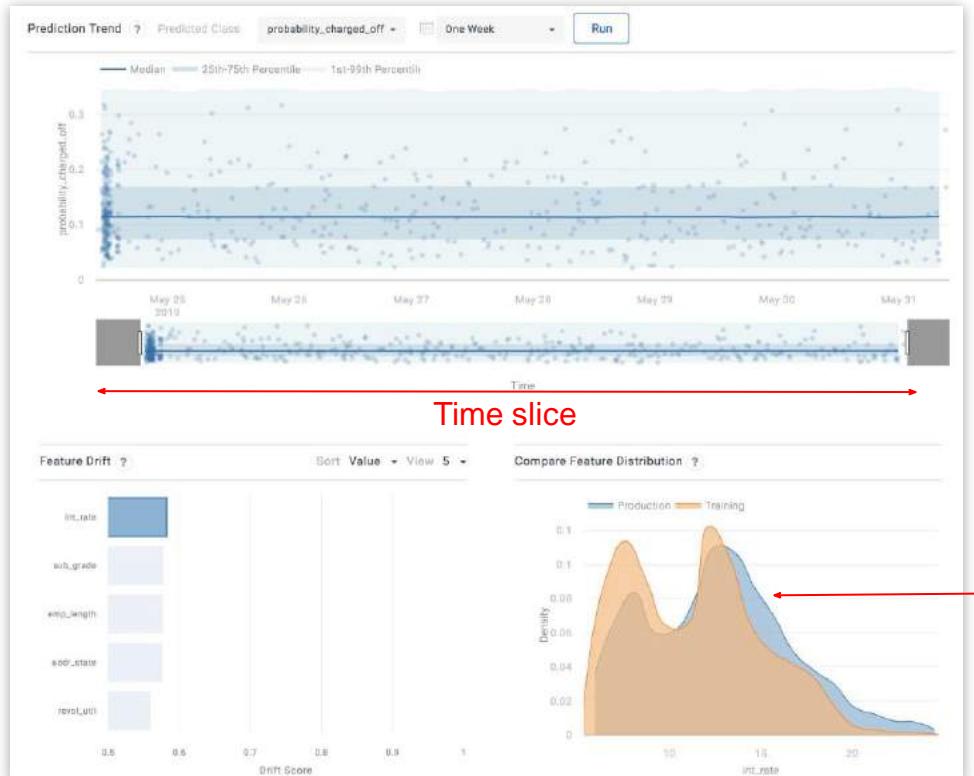
What are the primary feature drivers of the dataset on my model?

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

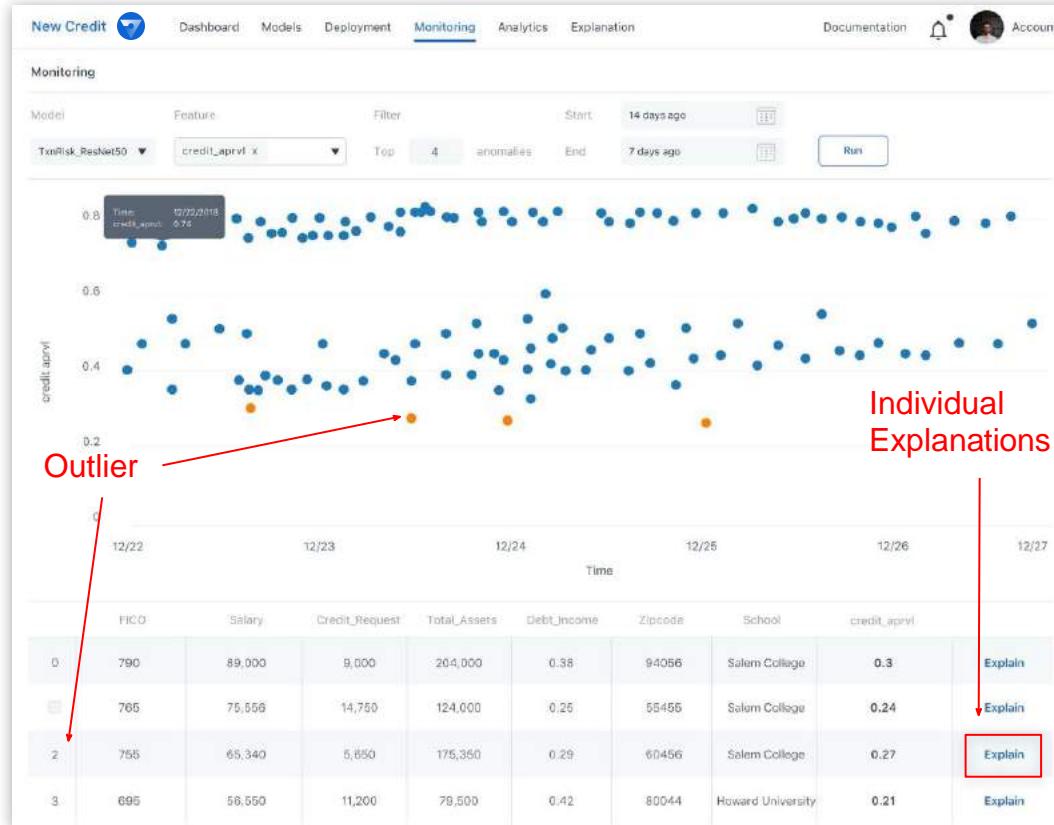


Investigate Data Drift Impacting Model Performance

Feature distribution for time slice relative to training distribution



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

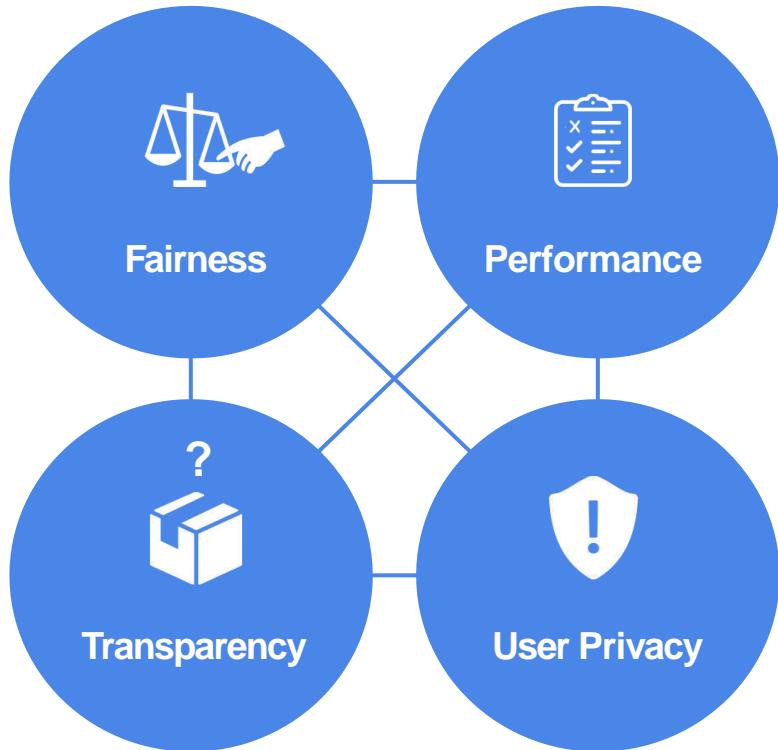
- Attributions are contrastive to their baselines
- Explaining explanations is important (e.g. good UI)
- In practice, we face engineering challenges as much as theoretical challenges

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