Explainable AI
come interpretare le predizioni di sistemi basati su AI e non solo

Tommaso Teofili

AI - Success stories

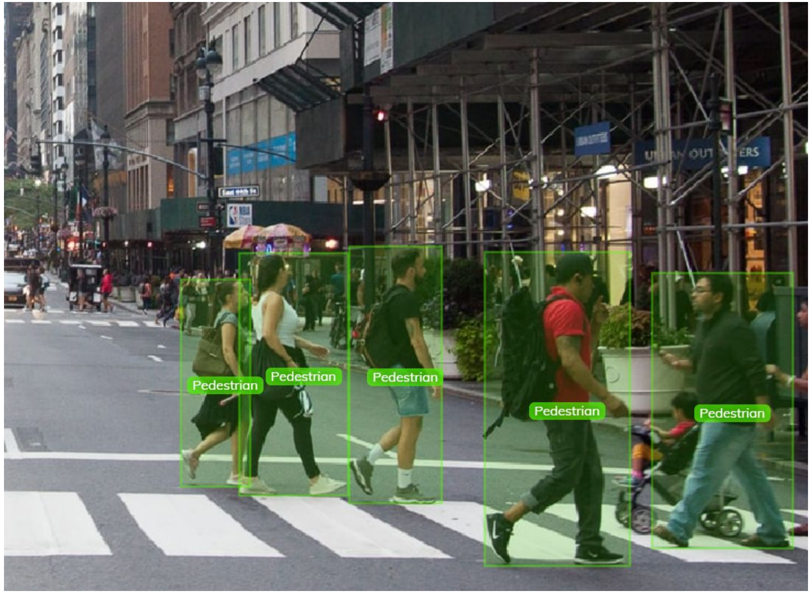
- Natural language

Datascienceseed provides **a list of resources that describe some of the most prominent technologies and applications in the area. It is a great place to start for an introduction to the topic of data science and machine learning.**

Written by Transformer · transformer.huggingface.co 🦄

AI - Success stories

- Vision



AI - Success stories

- Natural language
- Vision

TEXT DESCRIPTION

An astronaut Teddy bears A bowl of soup

riding a horse lounging in a tropical resort in space playing basketball with cats in space

as a children's book illustration in a minimalist style in a watercolor style

→

DALL·E 2

AI - Success stories

- Credit scoring

AI - Success stories

- Trading

AI - Success stories

- Data integration
- Building knowledge bases
- ...

OK, WE'RE DONE!

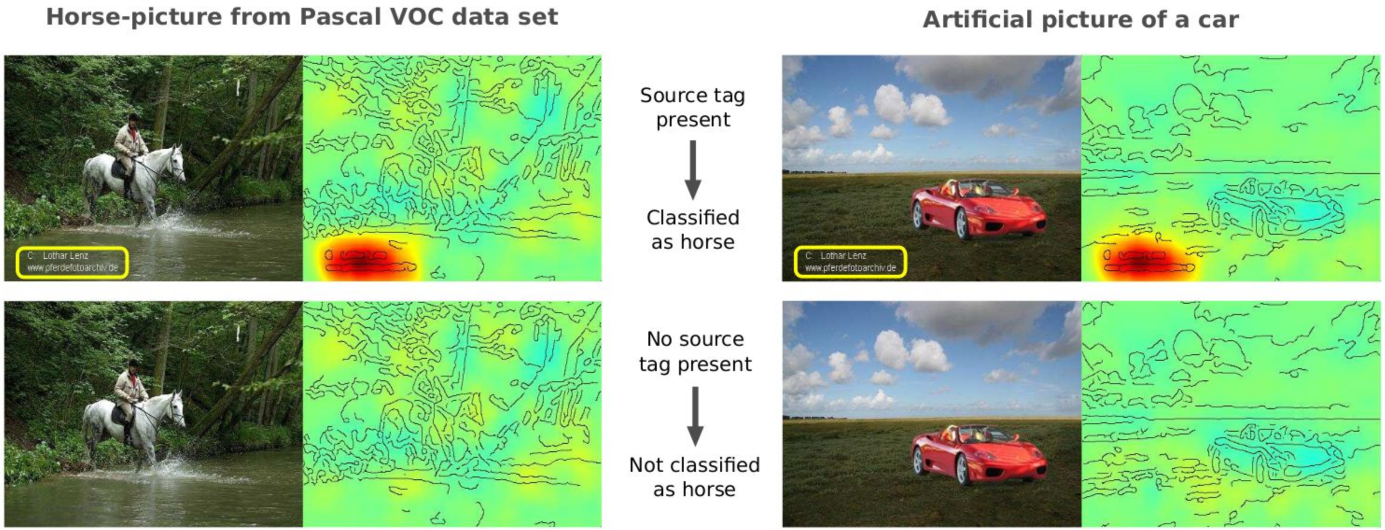
OK, WE'RE DONE!

AREN'T WE?

AI - Horror stories

- Recidivism risk prediction

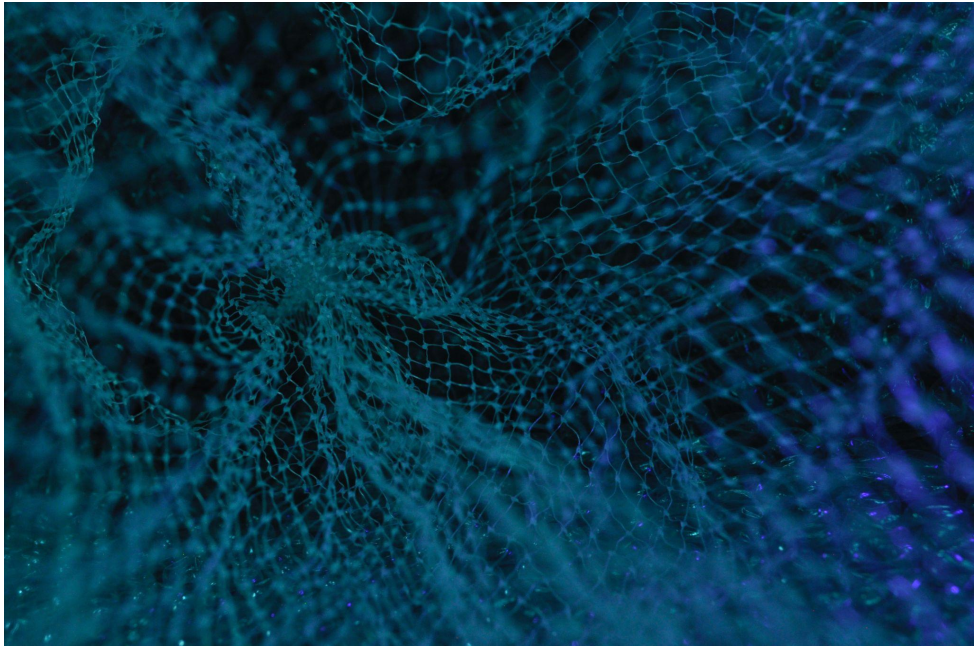
“We’re blind to the obvious” – D. Kahneman



“We’re blind to the obvious” – D. Kahneman

(a) Husky classified as wolf (b) Explanation

“We’re blind to our blindness” – D. Kahneman



ExplainableAI

- Interpretability via **explanations**
- An explanation is a **human understandable** description of an AI model / prediction internals

- ... ideally explanations expose human concepts in a sufficiently abstract (?) way so that they are easily understandable (?) by anyone (?) ...

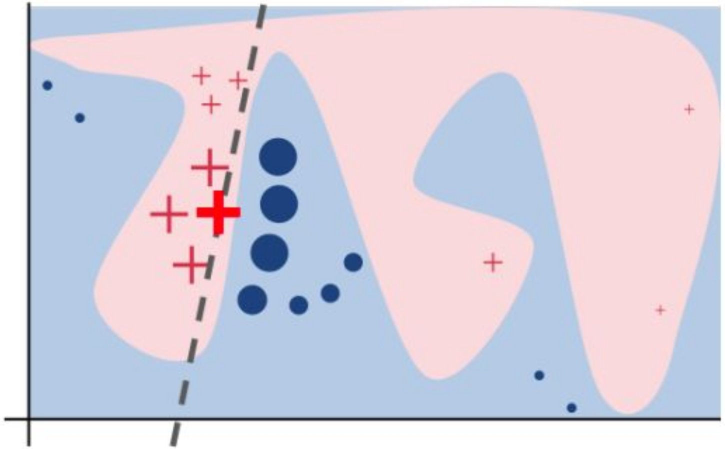
ExplainableAI

- “AI predicted that two DB records refer to the same real world entity. Why? Can I trust it?”
- “AI denied loan to applicant A while approved it for applicant B although their profiles look similar. Why? Can I trust it?”
- “AI did a code review on my pull request and rejected it. Why? What should I change to get it approved and merged?”

ExplainableAI
Generic black-box methods

LIME

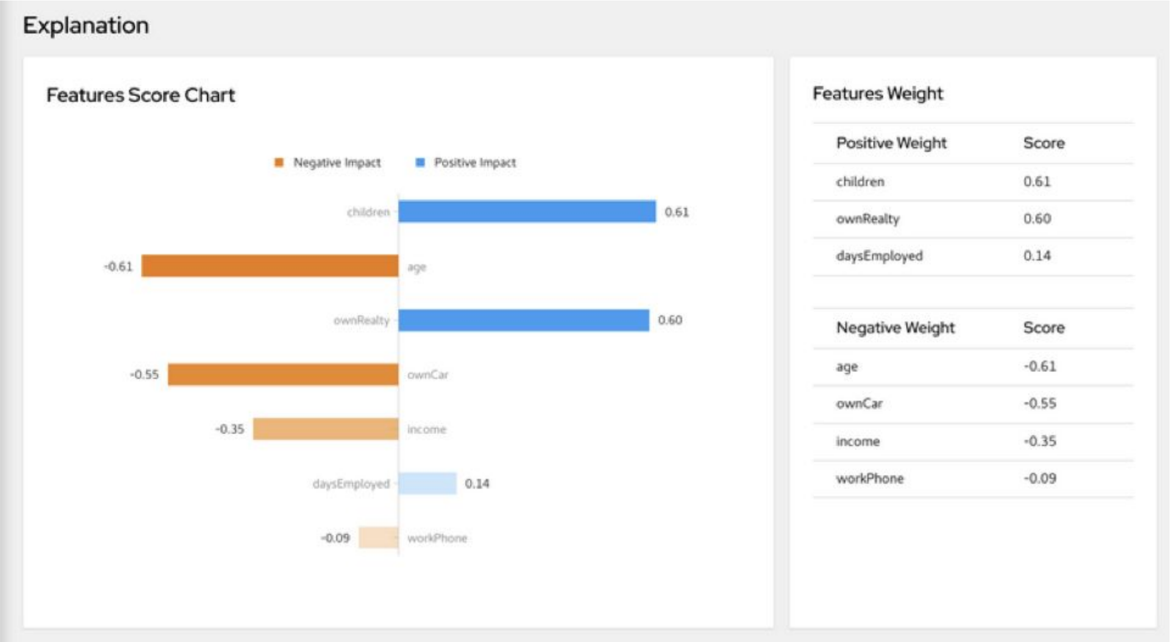
- Local Post-hoc method
- Trains an interpretable model in the "neighborhood" of the prediction input
- Generates an importance score (weight) for each feature in the input
- Model is treated as a black box



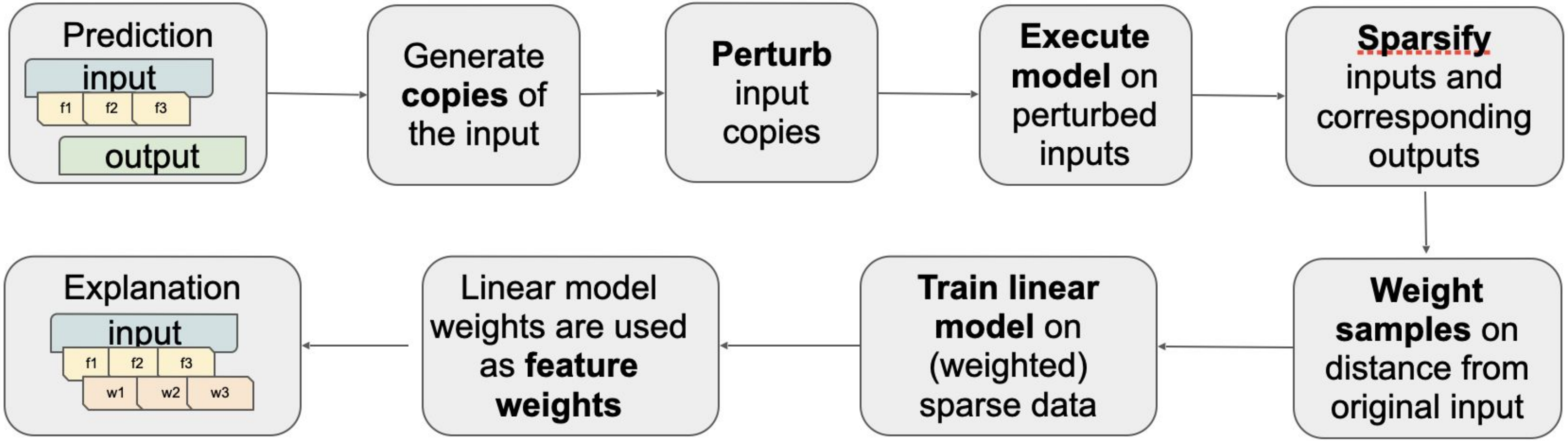
- Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. "*Why should i trust you?: Explaining the predictions of any classifier.*" Proceedings of the 22nd ACM SIGKDD, 2016.

ExplainableAI – Loan approval example

- Explanation for a positive loan approval prediction

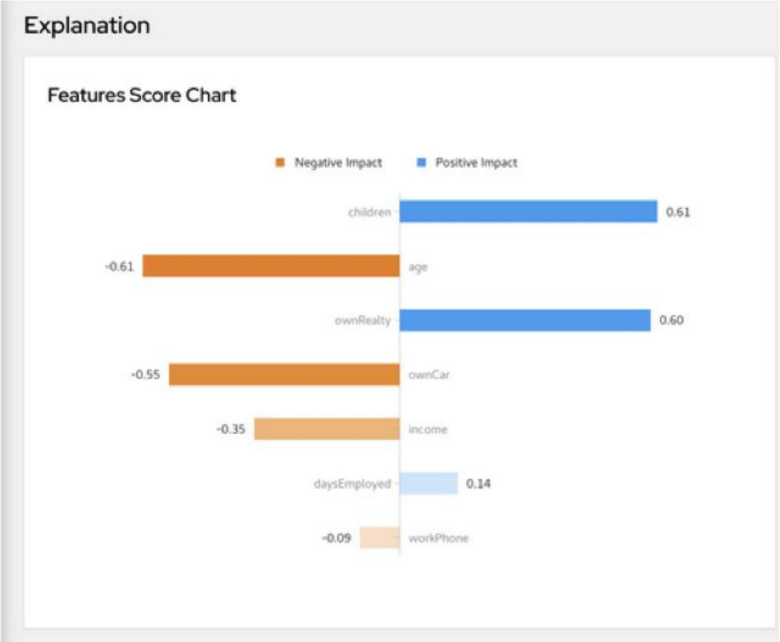


LIME - Workflow



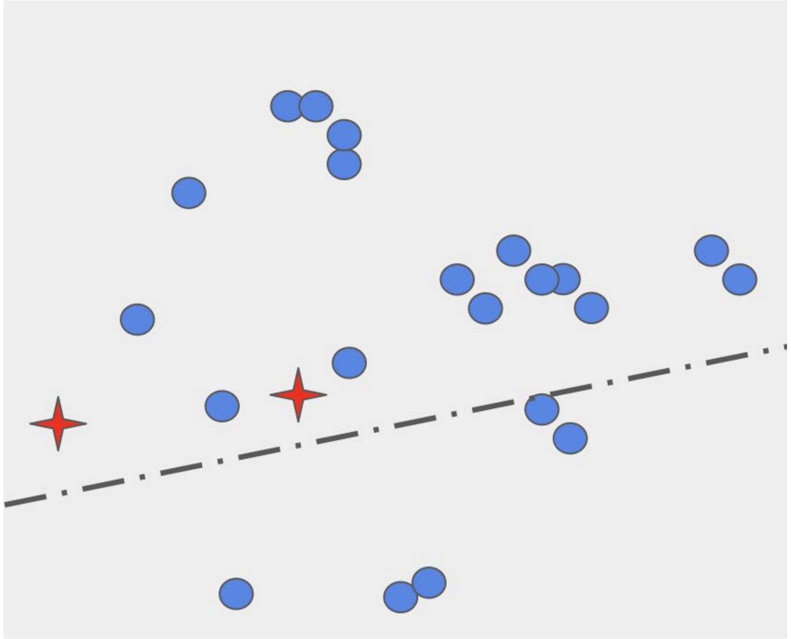
LIME

- Classification
  - Positive score for a feature means that feature is important for the model to predict true
  - Negative score for a feature means that feature is important for the model to predict false
- Regression
  - Positive score for a feature means that feature is important for the model to make that specific prediction value
  - Negative score for a feature means that feature is important for the model not to predict that specific value (a *contrastive* feature)
- Good general purpose starting point
- Issues
  - Stability
  - Sensitivity



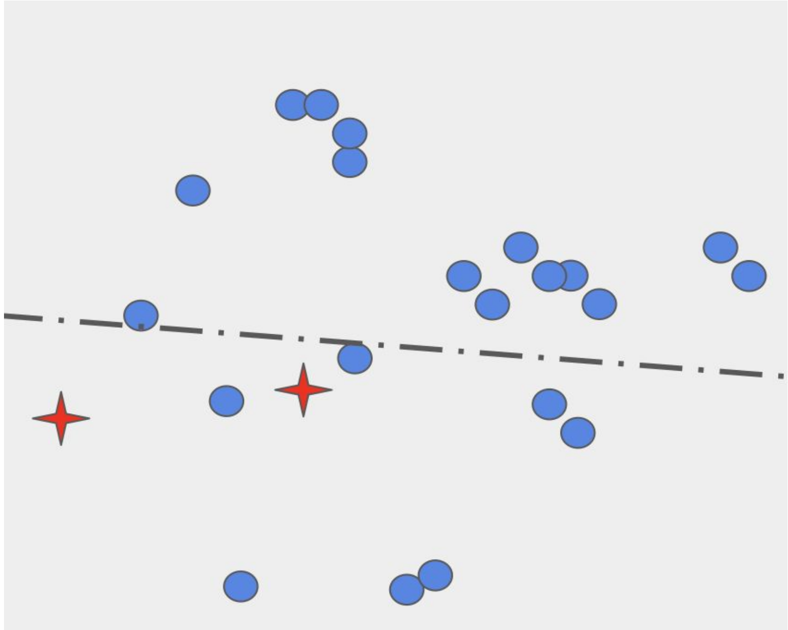
LIME – Can we do better ?

- Not enough samples to find a good decision boundary



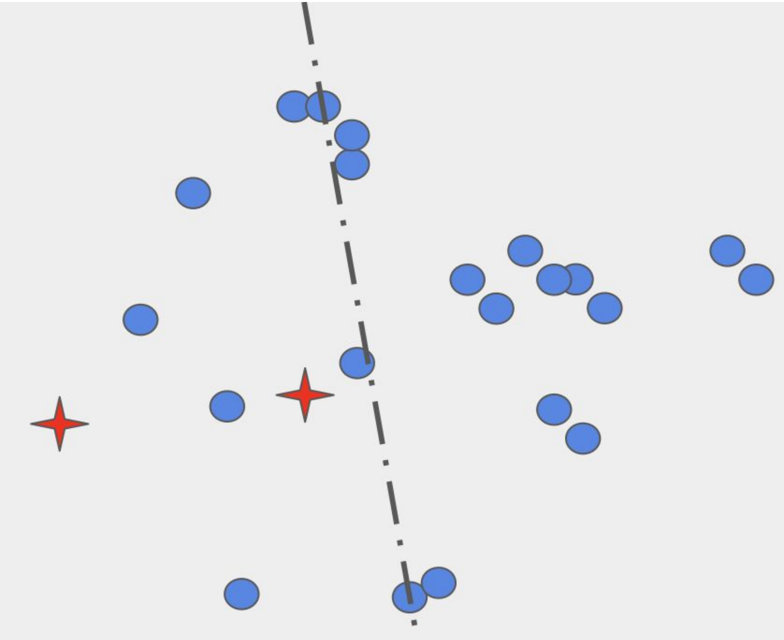
LIME – Can we do better ?

- Not enough samples to find a good decision boundary



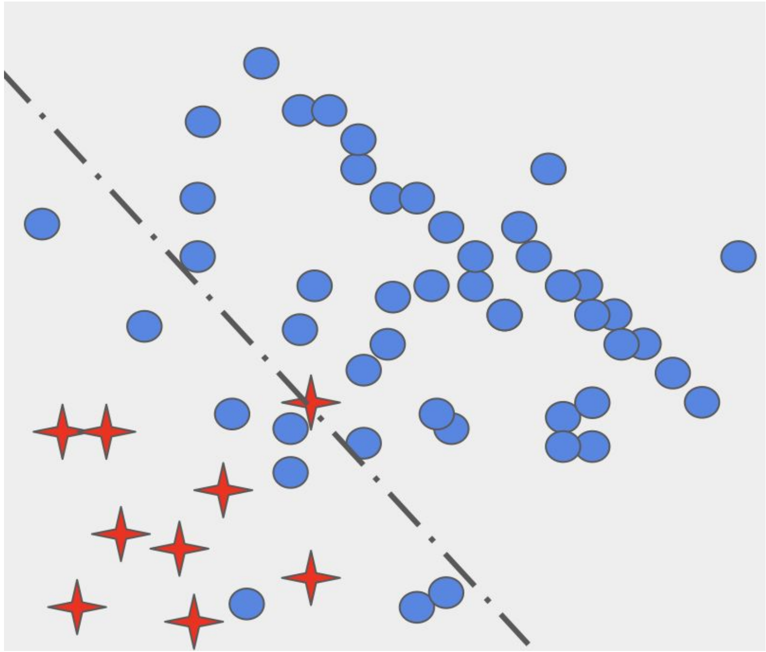
LIME – Can we do better ?

- Not enough samples to find a good decision boundary



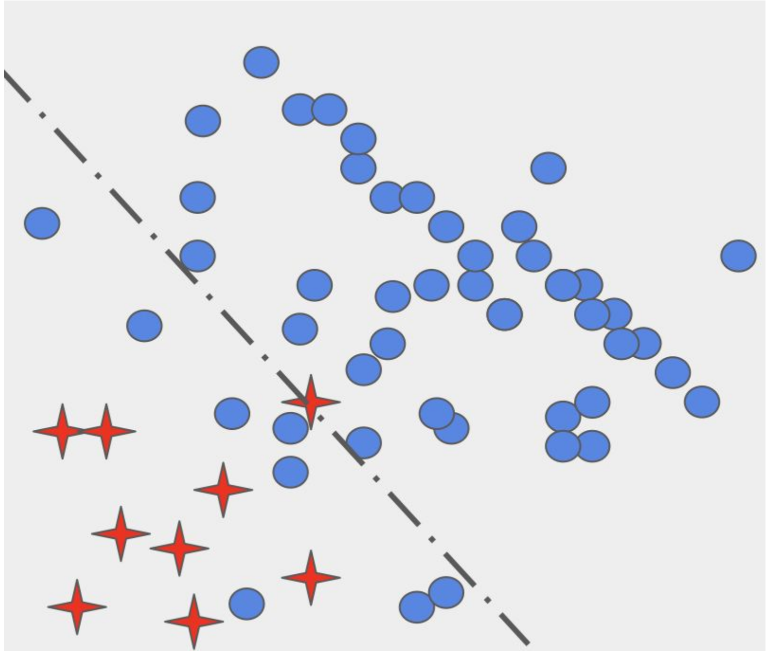
LIME – Adaptive Sampling

- Detect when the generated sparse dataset is highly unbalanced
  - E.g., 90% or more of the samples are predicted with the same class
- Generate more samples
- Increase the variance in the perturbation process
- Especially useful when the model has biased behaviors



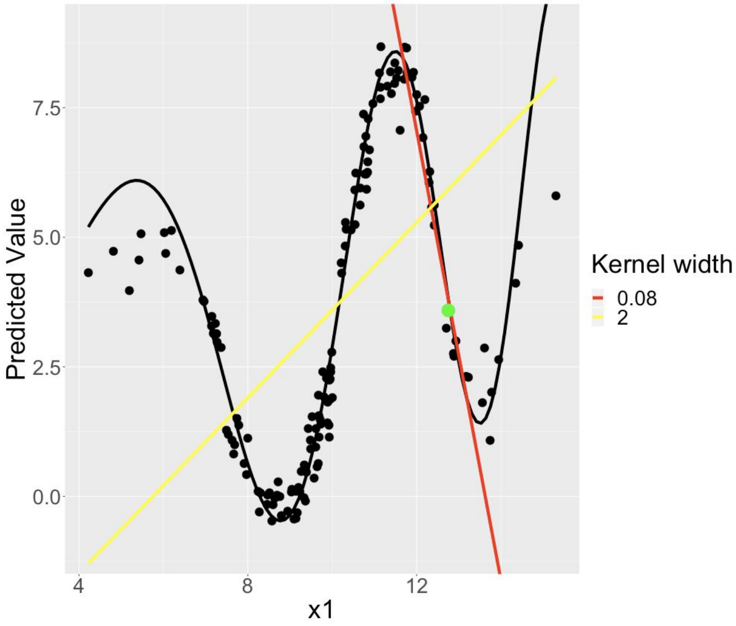
LIME – Adaptive Sampling

- **Adaptive sampling** allows to incrementally add samples as needed to fit a good decision boundary



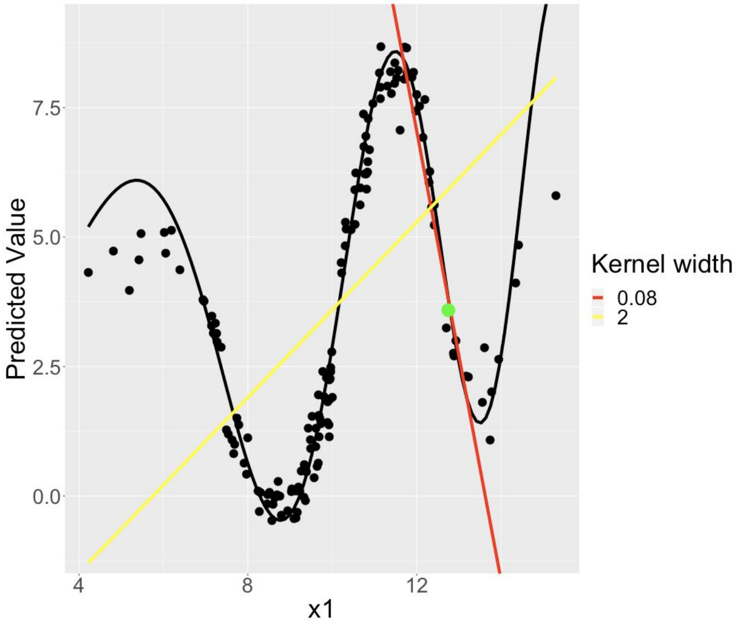
LIME – Can we do better ?

- Samples are assigned suboptimal proximity weights
- Might result in selecting wrong neighbors



LIME – Can we do better ?

- Filter samples (aggressively) rather than weigh them results to be more effective when combined with *adaptive sampling*



SHAP

- Relying on **Shapley value** from game theory as output importance scores
- Additive feature attribution
    - Helps understanding how the AI based system comes up with the output score for a given input
- More fine grained understanding of the system behavior
- Local and Global

Lundberg, Scott M., and Su-In Lee. *"A unified approach to interpreting model predictions."* Advances in neural information processing systems 30 (2017).

Shapley values

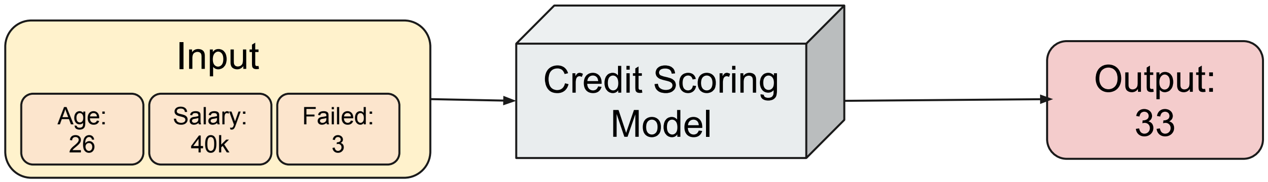
- If we have a **coalition** of **individuals** that collaborates to produce a certain **output**
- We want to know how much each individual contributed to it
- For each *individual*
    - We calculate the difference between the produced *outputs* when
        - The *individual* participates
        - The *individual* doesn't participate
        - For any possible *coalition* that includes that *individual*
    - The mean of these marginal contributions is the Shapley value

Shapley values → SHAP

- Individuals → Features
- Produced output → the AI system prediction
- Dropping Individuals from Coalitions → Selecting feature values from a *background dataset*

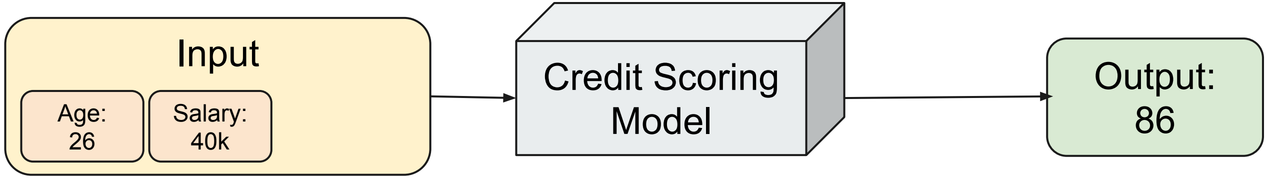
Shapley Value – Example

- Original prediction to be explained



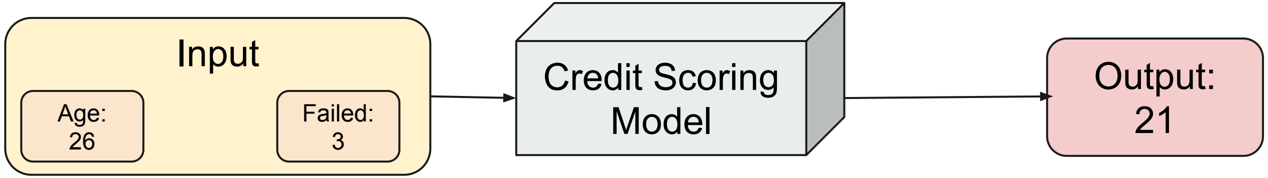
Shapley Value – Example

- Shapley Value for “Failed” feature
- Impact @ coalition #1 = 33-86



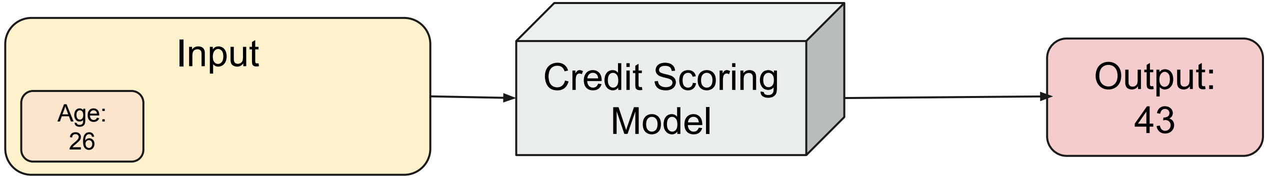
Shapley Value – Example

- Shapley Value for “Failed” feature
- Impact @ coalition #1 = 33-86



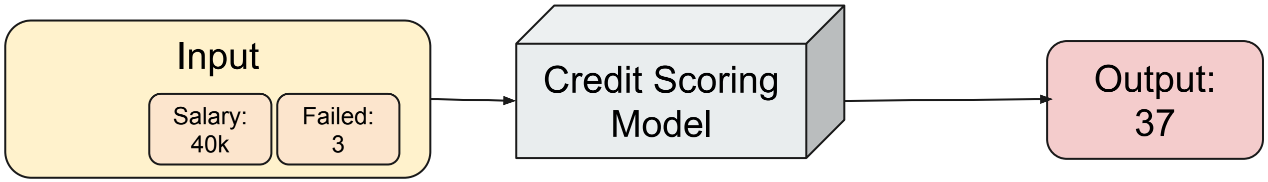
Shapley Value – Example

- Shapley Value for “Failed” feature
- Impact @ coalition #1 = 33-86
- Impact @ coalition #2 = 21-43



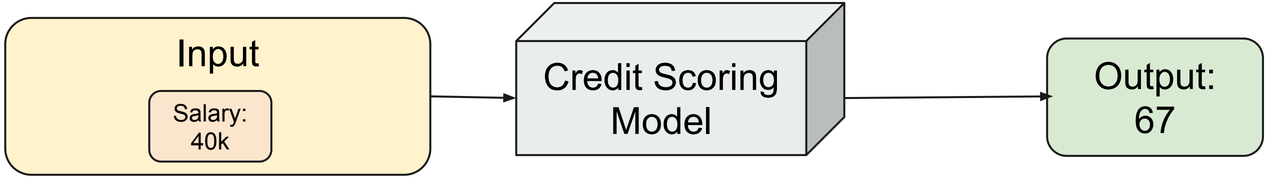
Shapley Value – Example

- Shapley Value for “Failed” feature
- Impact @ coalition #1 = 33-86
- Impact @ coalition #2 = 21-43



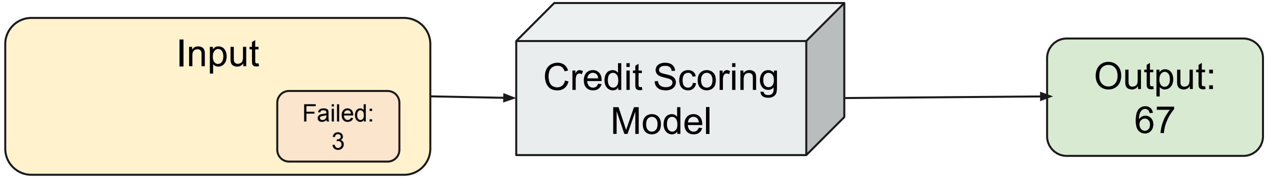
Shapley Value – Example

- Shapley Value for “Failed” feature
- Impact @ coalition #1 = 33-86
- Impact @ coalition #2 = 21-43
- Impact @ coalition #3 = 37-67



Shapley Value – Example

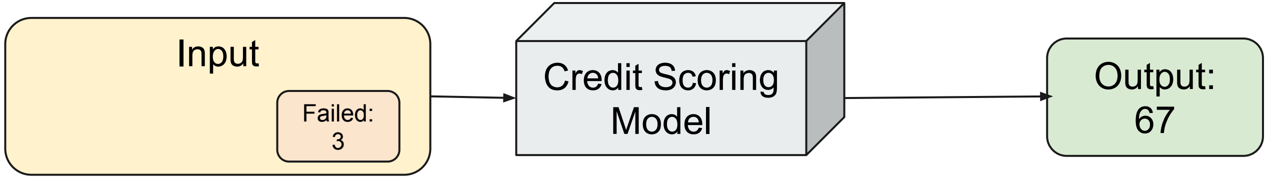
- Shapley Value for “Failed” feature
- Impact @ coalition #1 = 33-86
- Impact @ coalition #2 = 21-43
- Impact @ coalition #3 = 37-67



- Shapley Value for Failed is -35 → very **negative** impact on the score

SHAP – Background data

- Instead of not considering the "Failed" input feature
- The "Failed" feature gets a bunch of values from a **background dataset**



Shapley values → SHAP

- Compute Shapley value for each feature
  - Very expensive for more than 4-5 features
  - Approximate Shapley value calculation
    - Kernel SHAP and others

SHAP - Additivity

- Given an input $\{x_1 \ldots x_n\}$ predicted as y by a black box model
- **Kernel SHAP** learns a weighted linear regression that approximate Shapley values w_i for each feature such that
  - $\sum w_i * x_i = y$

SHAP - Can we do better ?

- The disadvantage of approximating **feature exclusion** via background data is that it renders all feature attributions as comparisons to the background data, not against **“true exclusion”**.
- If the background data selected is poor, the Shapley values will be less accurate
- What is a good “missingness” value for
  - “Failed” feature ?
    - Population mean numbers of failed payments ?
    - Zero ???
  - “Age” feature ?
    - Population mean age ?
    - Zero ???
  - ...

SHAP - Background data selection

- Pick samples that are
  - Similar to the original input
  - Similar to each other
  - Differently (evenly) predicted by the black box model
- Eventually let the user select a reference starting point

Counterfactuals explanation

- What should I change in my input for the black box model to predict a desired outcome, different than the actual one ?
- Explanation methods that generate new inputs
    - Close to the original
    - Predicted as desired
- E.g.,
    - *“My loan got rejected by the AI! What can I do ?”*
    - *“If you get a salary increase by 1000€ then you’ll get it”*
    - *“Or you should consider checking your bank account before attempting (and failing) payments”*
    - …

CPS based Counterfactuals

- Constraint Problem Solvers (CPS) are a family of algorithms that provide solutions by exploring a formally defined problem space (using **constraints**) to maximize a calculated **score**
- Let some features remain fixed
  - E.g., “Age”, I can’t get younger
  - E.g., “Number of children”, that doesn’t get changed quickly
- Define hard scores
  - Desired outcome met
  - Fixed features have not to be changed
- Soft scores
  - Distance between original and changed input features (e.g. Manhattan distance)

TrustyAI Explainability Toolkit

- Explainability tools
  - LIME*
  - SHAP*
  - CPS Counterfactuals
  - …
- Available both for Java and Python

https://arxiv.org/abs/2104.12717

https://github.com/trustyai-explainability/

TrustyAI Initiative

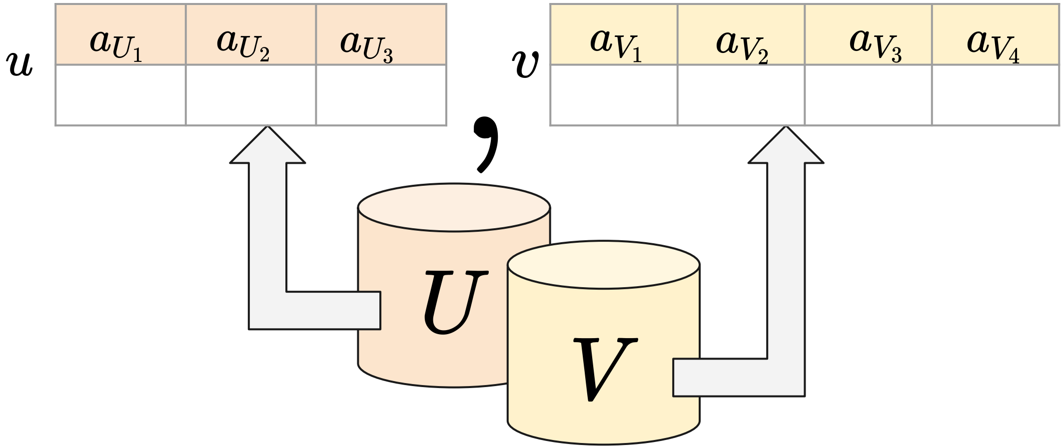
ExplainableAI
Principled solutions for specific tasks

Data integration

- Integrate **data** coming from different data sources
  - Clean
  - Normalize
  - Fix
- Continuous building of knowledge bases

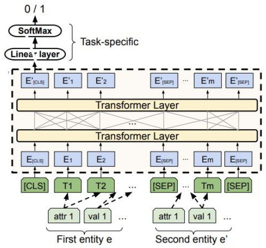
Entity Resolution

- Determine whether two **records** refer to the same **entity** in the real-world
- Part of the typical data integration pipeline

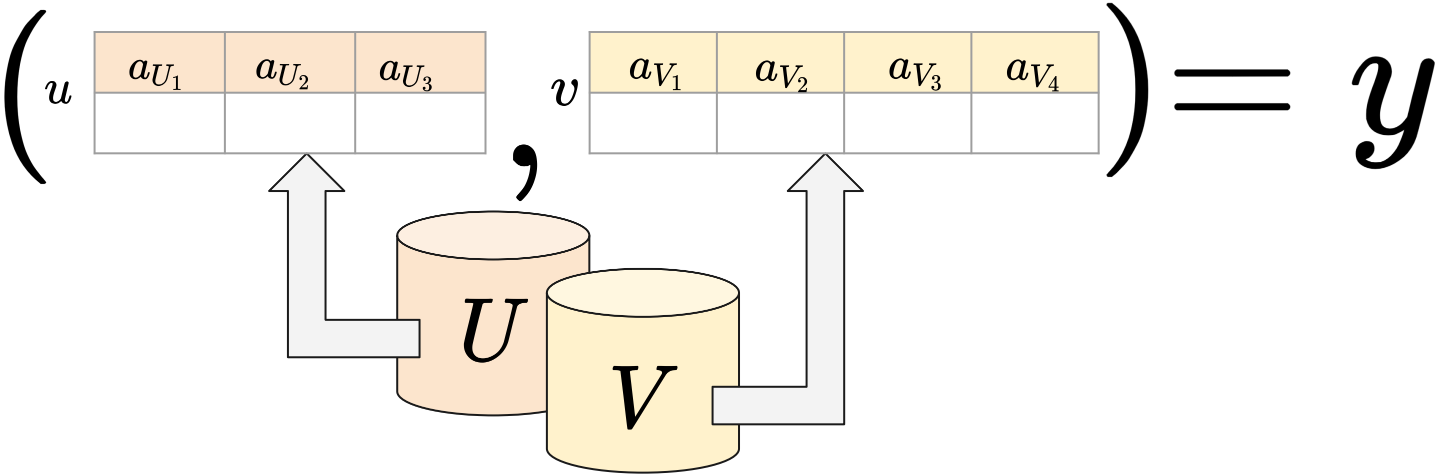


Solving Entity Resolution tasks with ML / DL

- Train a binary classification model M
- Predicts whether $\langle u, v \rangle$ records are **matching**

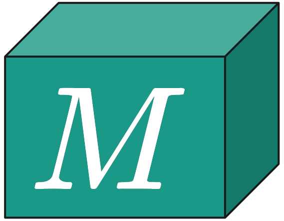


Ditto
[Li+; VLDB'20]



Solving Entity Resolution tasks with ML / DL

- Highly accurate



$$M\left(u \begin{array}{|c|c|c|}\hline a_{U_1} & a_{U_2} & a_{U_3} \\ \hline & & \\ \hline\end{array}, v \begin{array}{|c|c|c|c|}\hline a_{V_1} & a_{V_2} & a_{V_3} & a_{V_4} \\ \hline & & & \\ \hline\end{array}\right) = y$$

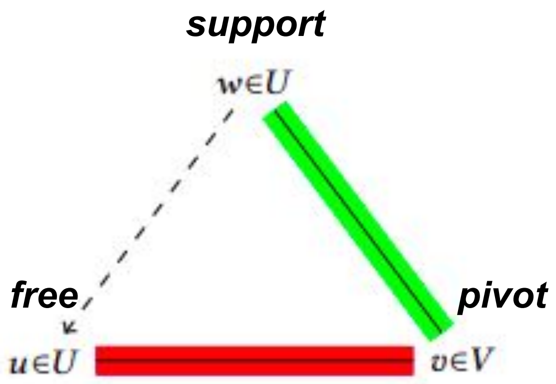
- Hardly interpretable
- No rationale for their predictions

Computing Entity Resolution explanations with TriAngles

- Principled framework for explaining ER models' predictions
- Attribute-level explanations
- Saliency explanations
  - Probability of necessity
- Counterfactual explanations
  - Probability of sufficiency

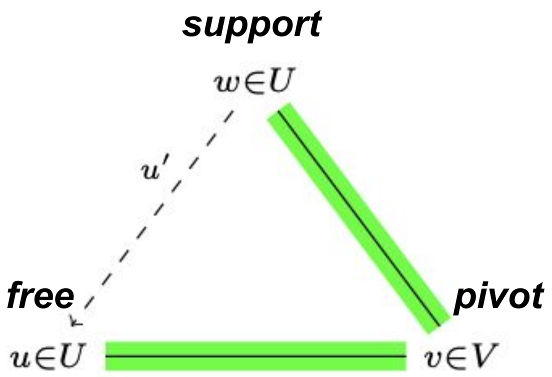
Teofili, Tommaso, et al. "Effective Explanations for Entity Resolution Models." arXiv preprint arXiv:2203.12978 (2022).

Open Triangle



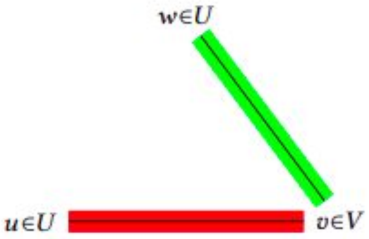
(a) $M(\langle u, v\rangle) = \mathbb{F}, M(\langle w, v\rangle) = \mathbb{T}$

Open Triangle Perturbation

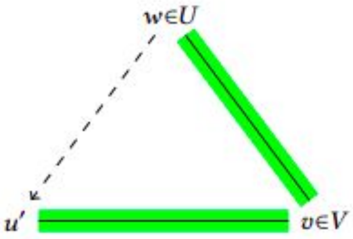


(b) Making the perturbed version of u, denoted u', more similar to w by copying values from w to u triggers $M(\langle u', v \rangle) = \mathbb{T}$.

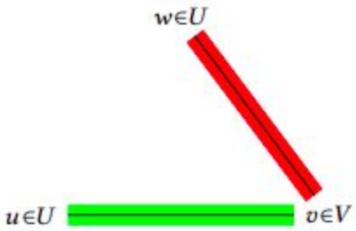
Open Triangles



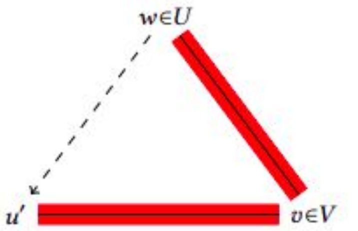
(a) $M(\langle u, v\rangle) = \mathbb{F}, M(\langle w, v\rangle) = \mathbb{T}$



(b) Making the perturbed version of u, denoted u', more similar to w by copying values from w to u triggers $M(\langle u', v\rangle) = \mathbb{T}$.



(a) $M(\langle u, v\rangle) = \mathbb{T}, M(\langle w, v\rangle) = \mathbb{F}$



(b) Making the perturbed version of u, denoted u', more similar to w by copying values from w to u triggers $M(\langle u', v\rangle) = \mathbb{F}$.

Probability of Necessity – Saliency

- The saliency of each attribute is calculated as the probability that the open triangle perturbation operation alters that attribute, conditioned to the fact that the prediction flips

$$\phi_a = P(u' \in \mathcal{U}_a | M(\langle u', v \rangle) = \overline{y})$$

[Watson+; UAI'21]

$$PN(c, y) := P(c(z) = 1 \mid f(z) = y).$$

Probability of Sufficiency – Counterfactuals

- A counterfactual explanation for $M(\langle u, v\rangle) = y$ is:
- A pair of records $\langle u', v'\rangle$ whose changed attributes A have the highest probability of sufficiency that changing them yields a prediction flip, with A being as small as possible

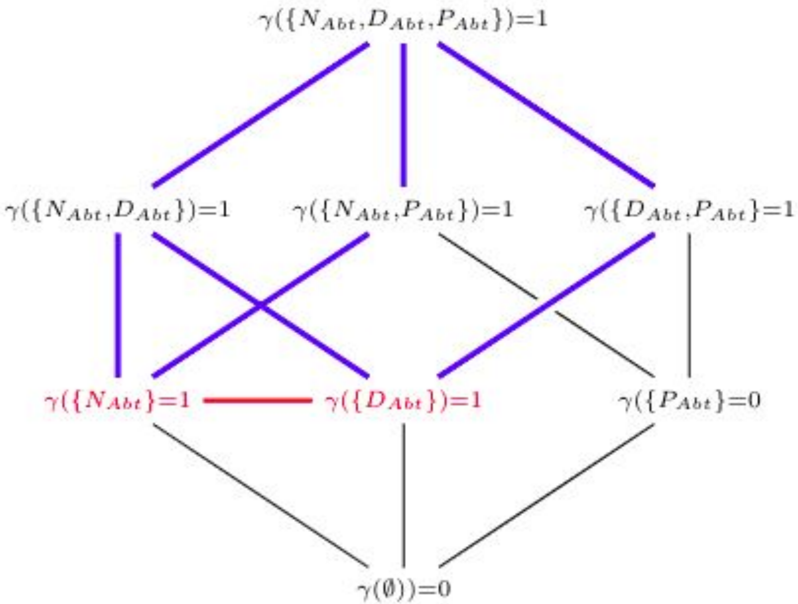
$$A^{\star} = argmin_A(|argmax_{A \subset \mathcal{P}(A_U) \setminus A_U} \chi_A|)$$

with $\chi_A = P(M(\langle u', v\rangle) = \bar{y} | u' \in \mathcal{U}_A)$ [Watson+; UAI'21]

$$PS(c, y) := P(f(\boldsymbol{z}) = y \mid c(\boldsymbol{z}) = 1).$$

Computing Probabilities on Lattice structures

- Starting from the bottom
- Proceed breadth-first
- At each node:
  - Perturb the corresponding attributes
  - If prediction flips, stop
  - Assume it flips also when perturbing supersets of attributes

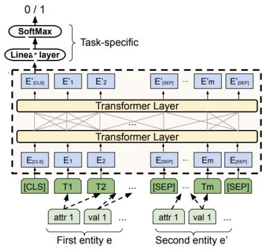


- *Monotonic classifier* assumption

[Tao; PODS'18]

CERTA - Example

- Help identify the *rationale* behind a classifier's predicted outcome



(

Name Abt	**Description** Abt	**Price** Abt
altec lansing inmotion portable audio system …	altec lansing inmotion ipod portable audio system im500usb…	NaN

,

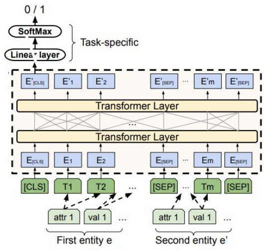
Name Buy	**Description** Buy	**Price** Buy
altec lansing inmotion im600 portable audio…		NaN

) $= 0.93$

- Is **Ditto** correctly predicting $\langle u, v\rangle$ as matching for sound reasons ?

CERTA – Example

- Help identify the *rationale* behind a classifier's predicted outcome



Name Abt	Description Abt	Price Abt
altec lansing inmotion portable audio system …	altec lansing inmotion ipod portable audio system im500usb…	NaN

Name Buy	Description Buy	Price Buy
altec lansing inmotion im600 portable audio…		NaN

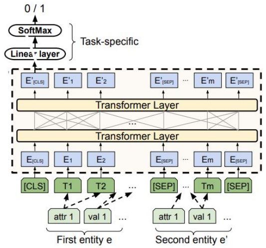
$= 0.93$

- Is **Ditto** correctly predicting $\langle u, v \rangle$ as matching for sound reasons ?

$$\Phi = \{\boxed{0.42, 0.43}, 0.27, \boxed{0.59}, 0.23, 0.33\}$$

CERTA – Example

- Help identify the *rationale* behind a classifier’s predicted outcome



(

Name Abt	Description Abt	Price Abt
sony 19 ‘ bravia m-series silver led flat panel hdtv …	sony 19 ‘ bravia m-series silver led flat panel hdtv …	NaN

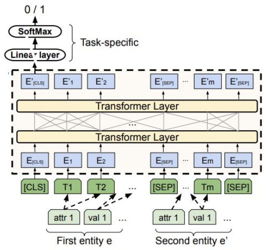
,

Name Buy	Description Buy	Price Buy
sony bravia m-series …	19 ‘ atsc 16:9 1440 x 900	379.72

) = 0.002

- Why is **Ditto** making this wrong non-matching prediction ?

CERTA – Example



Name Abt	Description Abt	Price Abt
sony 19 ' bravia m-series silver led flat panel hdtv …	sony 19 ' bravia m-series silver led flat panel hdtv …	NaN

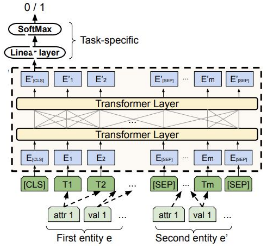
Name Buy	Description Buy	Price Buy
sony bravia m-series …	19 ' atsc 16:9 1440 x 900	379.72

$$= 0.002$$

$$\Phi = \{0.2, 0.11, 0.04, \boxed{0.29, 0.28}, 0.01\}$$

- Name_{Buy} and $\text{Description}_{\text{Buy}}$ are the two most influential features for this prediction

CERTA – Example



Name Abt	Description Abt	Price Abt
sony 19 ' bravia m-series silver led flat panel hdtv …	sony 19 ' bravia m-series silver led flat panel hdtv …	NaN

,

Name Buy	Description Buy	Price Buy
sony bravia m-series …	19 ' atsc 16:9 1440 x 900	379.72

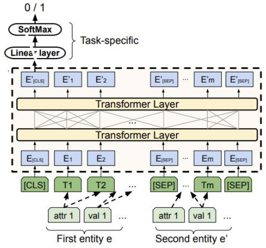
$= 0.002$

$$\Phi = \{0.2, 0.11, 0.04, \boxed{0.29, 0.28}, 0.01\}$$

- Name$_{Buy}$ and Description$_{Buy}$ are the two most influential features for this prediction

  → Their values only appear in negative samples in the training set!

CERTA – Example



(

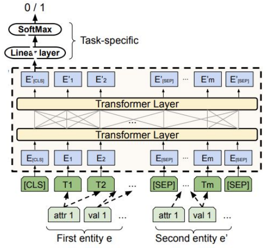
Name Abt	Description Abt	Price Abt
sony 19 ' bravia m-series silver led flat panel hdtv …	sony 19 ' bravia m-series silver led flat panel hdtv …	NaN

,

Name Buy	Description Buy	Price Buy
sony bravia m-series …	19 ' atsc 16:9 1440 x 900	379.72

) $= 0.002$

- A counterfactual explanation provides a new input $\langle u', v' \rangle$ that changes a prediction to a desired outcome



(

Name Abt	Description Abt	Price Abt
sony bravia system home …	sony 19 ' bravia m-series silver led flat panel hdtv …	NaN

Name Buy	Description Buy	Price Buy
sony bravia m-series …	19 ' atsc 16:9 1440 x 900	379.72

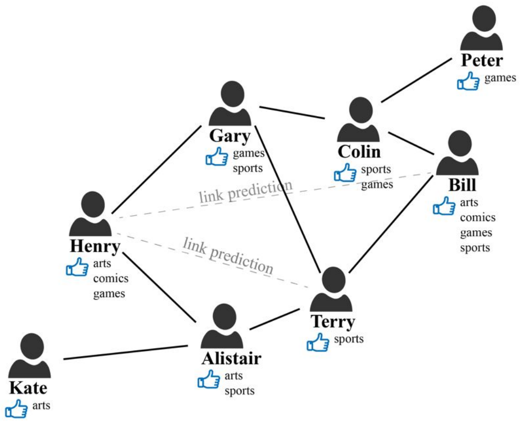
) $= 0.64$

Counterfactual data augmentation

- Pick counterfactual examples for misclassified inputs
- Add them (with caution) to the training data
- Retrain !

Link Prediction

- Another fundamental task in the context of data management
- Create, update and improve incomplete knowledge bases



Embeddings based Link Prediction models

- Many recent accurate models for LP are based on the notion of **embeddings**
- **Embeddings** are dense vectors that effectively represent “something” such that we can query similar such “something” in a **vector space**
- Usually “learned” by deep neural networks
    - E.g. Word2Vec, GloVe, ELMO, BERT, etc. in NLP
    - TransE, ComplexE, etc. in LP
- In LP they might represent Entities, Relations, Facts, etc.

Explain why certain ”links” are predicted

- Why is Barack Obama predicted as American? (correct)
- Why is Francesco Totti predicted as Major of Rome? (wrong)

Explain why certain ”links” are predicted

- Why is Barack Obama predicted as American? (correct)
- Why is Francesco Totti predicted as Major of Rome? (wrong?)

What are the most influential training facts for a given prediction ?

Rossi Andrea, et al. “Explaining Link Prediction Systems based on Knowledge Graph Embeddings”, ACM SIGMOD 2022.

Explainable AI – Conclusions

- Explainable AI techniques allow to gain better insights on the behaviors of complex opaque systems
- Explanations also provide ways to
  - Debug training data and learning procedures
  - Fix spurious patterns