XAI for Semantic Dependency

How to understand the impact of higher-level concepts on AI results

Holger Ziekow
Faculty of Business Information Systems
Furtwangen University
Germany
e-mail: holger.ziekow@hs-furtwangen.de

Peter Schanbacher
Faculty of Business Information Systems
Furtwangen University
Germany
e-mail: peter.schanbacher@hs-furtwangen.de

Presenter: Holger Ziekow

- Professor at Furtwangen University
  (Faculty of Business Information Systems)

- Research interest
  - Data Science and Machine Learning
  - XAI
  - Big and Streaming data
  - *Application areas:*
    (Manufacturing, Medicine, IoT, …)

The General Problem

- Inner workings of *black box* machine learning models are hard to understand
- Humans seek insights into model decisions

Solution:
XAI methods to analyze models

Image based on output from Stable Diffusion SDXL

XAI and Semantic Dependency Analysis as Solution

- XAI methods analyze black box models
- Our works introduces an XAI method to analyze how higher level concepts impact model decisions (semantic dependency)

XAI for Semantic Dependency

How to understand the impact of higher-level concepts on AI results

Holger Ziekow
Faculty of Business Information Systems
Furtwangen University
Germany
e-mail: holger.ziekow@hs-furtwangen.de

Peter Schanbacher
Faculty of Business Information Systems
Furtwangen University
Germany
e-mail: peter.schanbacher@hs-furtwangen.de

***Abstract*—XAI methods such as partial dependency plots or individual conditional expectation plots help understanding the impact of feature values on the output of an AI model. However, these techniques can only analyze the concepts manifested in a single feature. This makes it hard to investigate the impact of higher-level concepts, spanning across multiple features (E.g. a model prediction may depend on the morbidity of a patient, while morbidity is only indirectly reflected through features about symptoms). In this paper we present and test a concept for getting insight into model dependency on aspects on a higher semantic level. This enables an understanding how a model output changes in dependence on meaningful higher-level concepts and aids data scientists in analyzing machine learning models.**

Keywords-Interpretability, Understandability; Explainability; explainable AI; XAI; human-centered AI; black-box models

I. Introduction

Due to increasing computational power, improving algorithms and access to big-data, Artificial Intelligence (AI) models gained popularity in recent years. Applications range from healthcare (Lee et al., [15]; Chen et al., [6]), credit risk (Szepannek and Lübke, [23]), autonomous driving (Grigorescu et al., [14]; Feng et al., [9]), image classifications (Sahba et al., [20]), audio processing (Panwar et al, [19]), among others.

The large number of parameters and complex interactions makes most AI models (in particular deep neural networks) hard to understand and difficult to interpret the results. For many applications it is required not only to have a model with high accuracy but also explain the outcomes. Regulators (European Commission, [8]) require the understandability of these models, in particular to increase their trust (Lui and Lamb, [17]) and assess potential biases (Challen et al., [5]).

What "explainability" means is not well defined and might be misleading (Rudin, [27]). It further depends on the context of the application. For MRI scan, the explanation might be a heat map of relevant areas for the model. For sentiment analysis of user feedback, the explanation might be relevant words of the text. Surrogate models such as decision trees may give an insight into more complex models.

In general, explainability methods can be distinguished into either global explainability on the model level such as variable importance (Breiman, [4]), partial dependency plots (PDP, Friedman, [10]), or accumulated local effects (ALE, Apley and Zhu, [1]), or local explainability on the level of individual predictions such as Shapley values (SHAP, Shapley, [21] or Strumbelj and Kononenko, [22]), or local interpretable model explanations (LIME, Ribeiro et al., [24]).

We lean on the notion of partial dependency plots (PDP). However, unlike PDPs, we capture the dependency on a higher-level concept, and not a single feature. (E.g. a concept that manifests in many features or the combination of many feature values). The analysis shows the model output if a certain concept is more or less present. E.g. one may analyze if a medical model leans more or less towards a certain recommendation, dependent on the morbidity of a patient. Yet, the morbidity may not be an explicit input of the model but indirectly reflected in a set of features about certain symptoms. Another example is an image classifier. Existing methods analyze the impact of pixels or regions in specific figures (see e.g. Bulat & Tzimiropoulos, [3]). However, reasoning about the semantics of these regions is up to the analyst and must be done instance by instance. With our method, one gains an understanding how presence of a certain concept impacts the model output. To the best of our knowledge, this constitutes a new approach. In this context, we refer to the approach as semantic dependency analysis (not to be confused with semantic dependency in NLP). As an illustrative example, we analyze how the presence of vegetation impacts the classification of an image as showing a city or rural area.

Our main contributions are the following

- We present a new general concept which we call semantic dependency analysis (SDA).
- We provide formalisms to define two fundamental ways of implementing SDA.
- We describe a specific implementation along a sample case.
- We present experimental results that demonstrate the working and utility of the approach.

The remainder of the paper is structured as follows: The introduction is followed by section 2 presenting the current state of literature and how our approach fits into the related work. Section 3 defines the concept of Sematic dependency analysis (SDA) and presents a possible implementation for generators as well as prediction models. Section 4 shows how

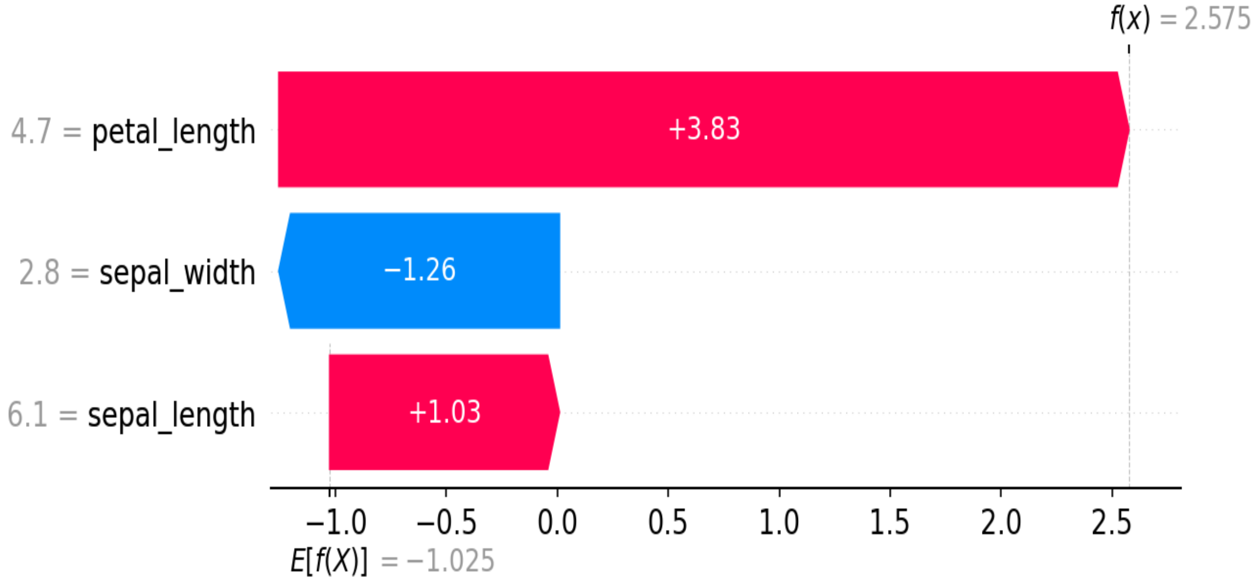
Outline

- Background XAI and Partial Dependency Plots
- Our Extension: Semantic Partial Dependency Analysis (SDA)
    - ***Implementation with generators***
    - ***Implementation with prediction models***
- Experiments with Sample Implementation
- Conclusion and Future Work

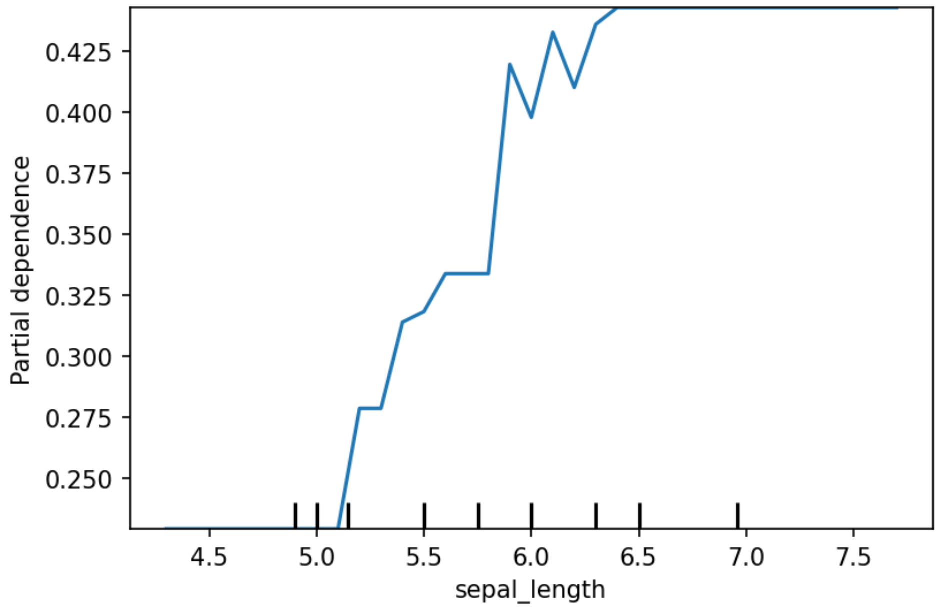
XAI and Partial Dependency Plots

- Many methods exist to analyze the effect of individual features on a black box model
- Examples include SHAP and partial dependency plots (PDP)

SHAP values for predicting "versicolor"



PDP for impact on sepal_length on predicting "versicolor"



Limitations of Partial Dependency Plots

- Analysis is limited of one feature at time[1]
- Impact of higher level concepts (not directly reflected in a feature) cannot be visualized

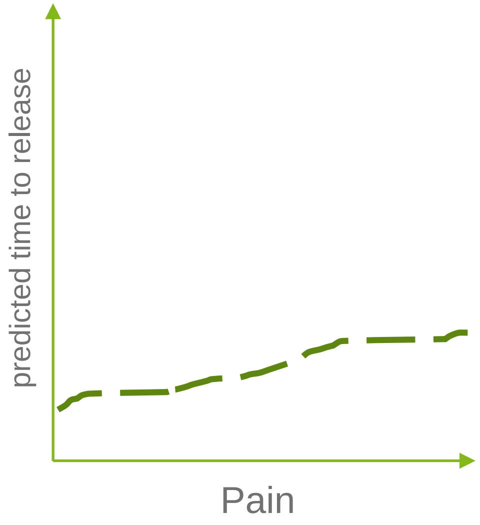
Illustrative artificial example

Morbidity

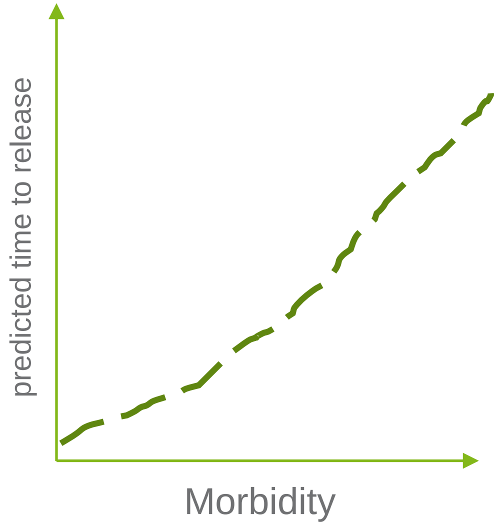
relates to

Appetite	Pain	Attitude	Weight	Time to release
3	1	3	10	7
1	8	1	12	10
5	1	5	14	11

Partial dependency on single feature



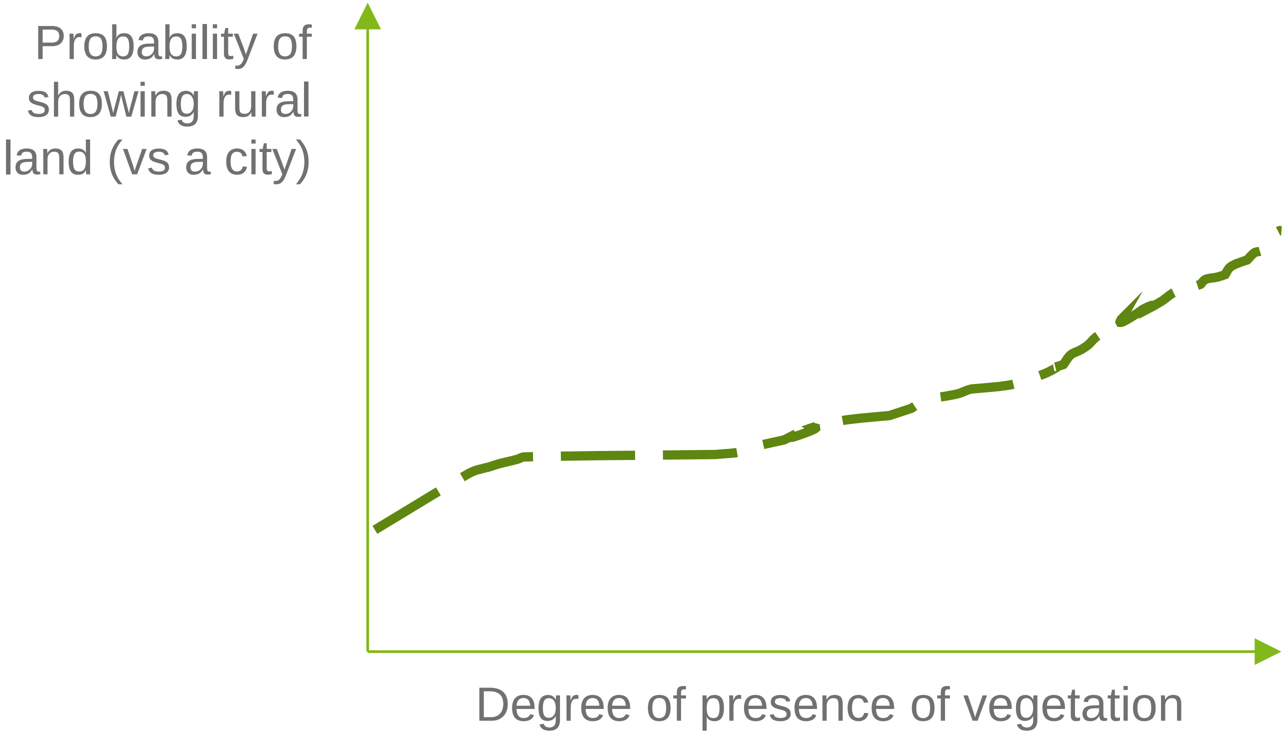
Partial dependency on higher level concept



[1]or a small set of features

Another example for higher level concepts

- Consider an image classification task for landscapes
- Analyze how the presence of vegetation impacts the model outcome
  - *Note: "presence of vegetation" is not a feature in the input data*



Semantic Dependency Analysis

Semantic Dependency Analysis (SDA)

- Idea: compute the expected model output for data instances that have the higher level concept present to the defined degree x_H

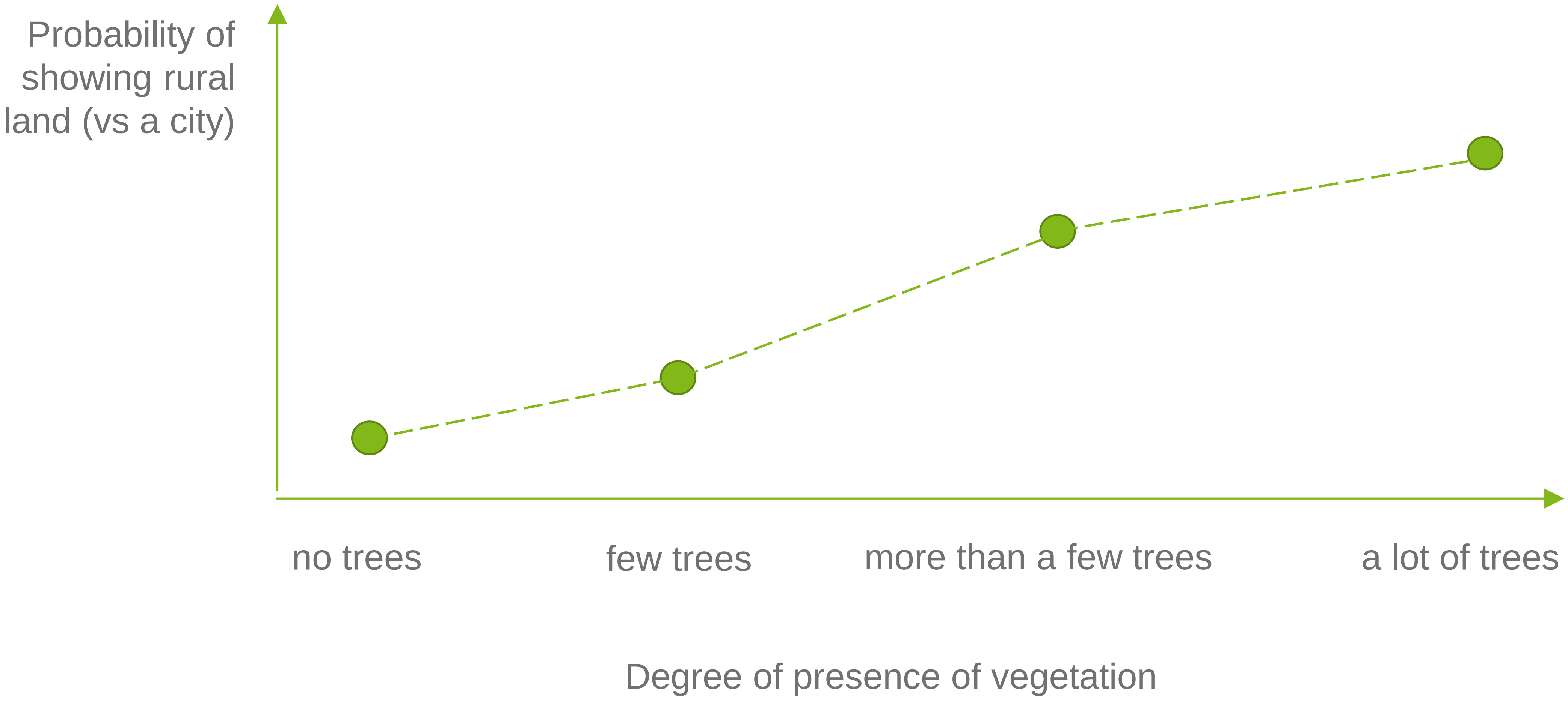
- Generate samples from the modeled domain X that have the concept to the specified degree x_H. (E.g. how much vegetation is present in a landscape image.)

ML model

$$SD_H(x_H) = E_X[\hat{f}(g(x_H, X))]$$

Random variable that returns feature vectors according to x_H and X

Illustrative Example



Implementing Semantic Dependency Analysis

- How to implement $\boldsymbol{g}(\boldsymbol{x}_{\boldsymbol{H}}, \boldsymbol{X})$?

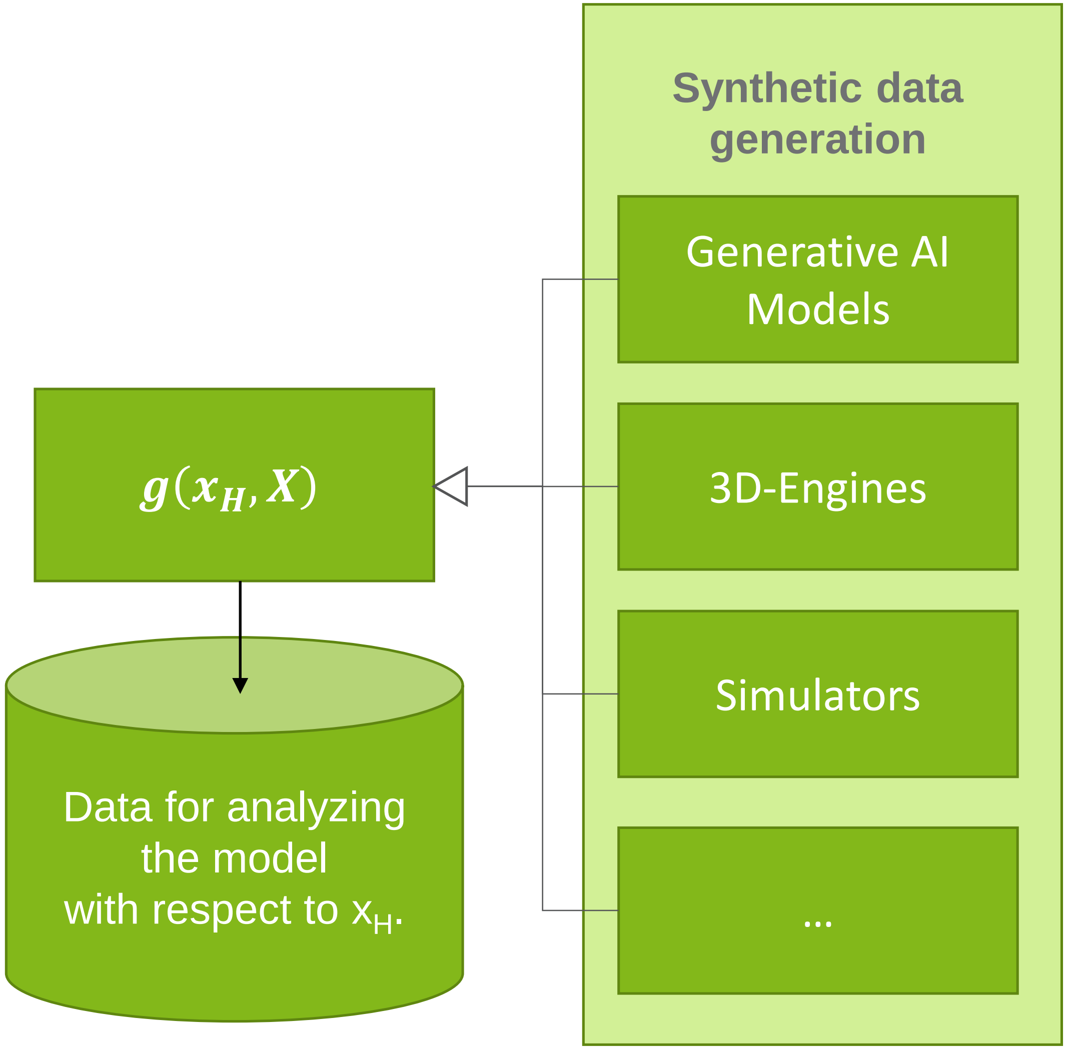
- Proposed methods

  - Implementation with generators

  - Implementation with prediction models

Implementating *g* with generators

Idea: Create synthetic data according to X and ensure that the analyzed concept is present to degree x_H.

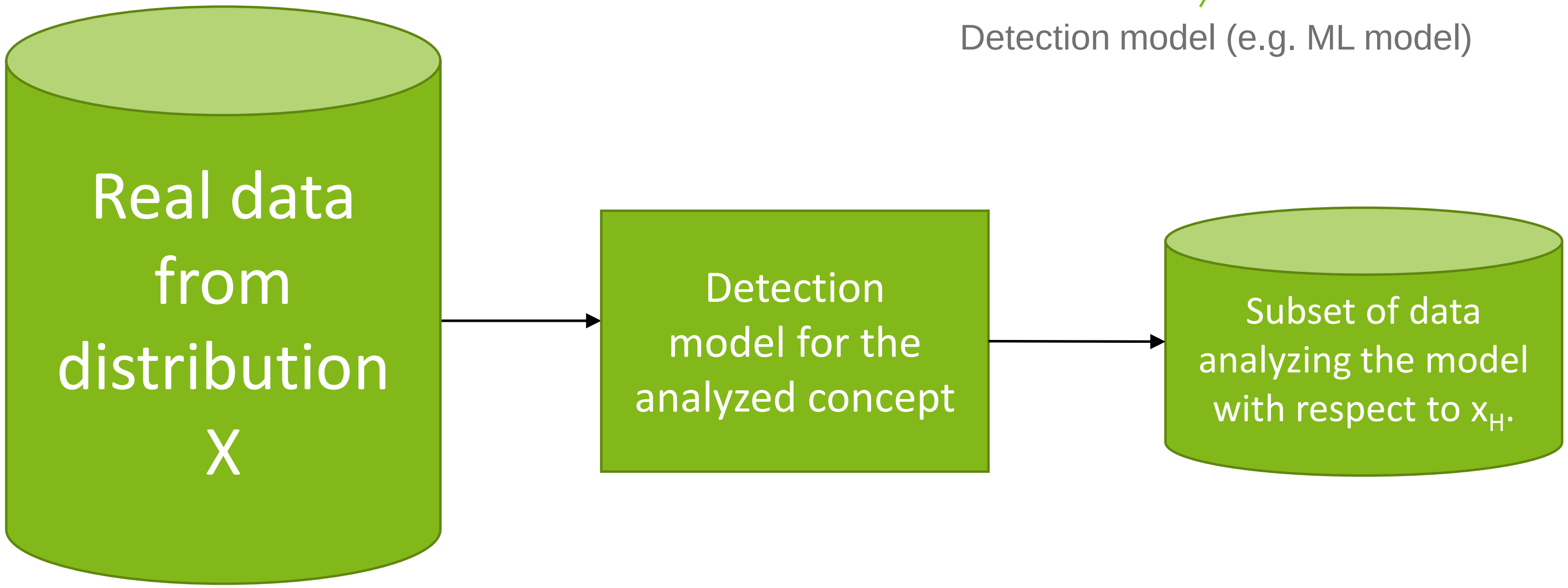


Implementating *g* with prediction models

Idea: Use real data from distribution X and filter out samples that have the analyzed concept is present to degree x_H.

$$SD_H(x_H, s) = E_X\left[\hat{f}(\{x \mid d(x, x_H) \in [s-\varepsilon, s+\varepsilon]\})\right]$$

Detection model (e.g. ML model)



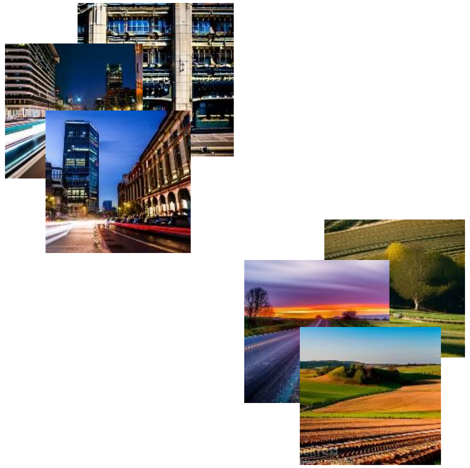
Experiments with Sample Implementation

Experimental Setup

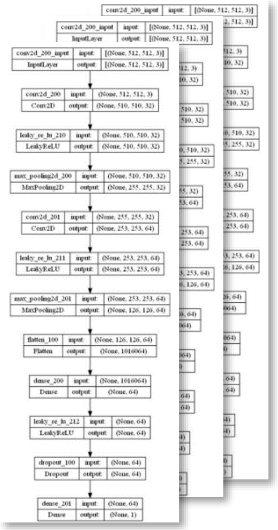
Image classification task as example

- Classify landscapes in “city” or “rural”
- Analyze the impact of the presence of trees on the model output

Step 1. Generating training set with images for classes “city” and “rural landscape” (using stable diffusion 2.0)



Step 2. Training multiple image classifiers for analysis

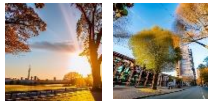


Arbitrary model generated by chatGPT and trained multiple times

Step 3. Generating images sets where the higher level concept “trees” is present to various degrees (using stable diffusion 2.0).

City with no trees

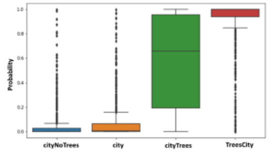
City with normal presence of trees

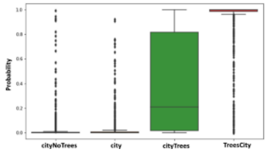


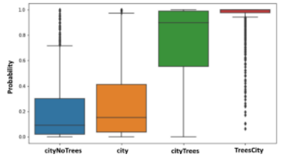
City with above normal presence of trees

City with very high presence of trees

Step 4. Conducting SDA with models form step 2 and images from step 3.







Generating Data for sample Classification Task

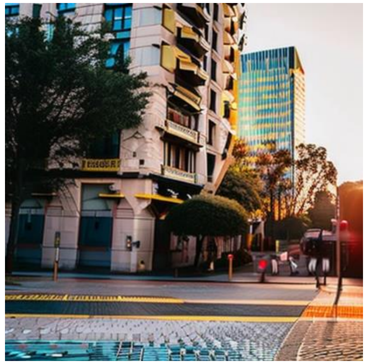
Using Stable Diffusion 2.0 with positive and negative prompts

Class “city”

Positive prompt: Photograph a city, high quality photography, Canon EOS R3

Negative prompt: digital art, drawing

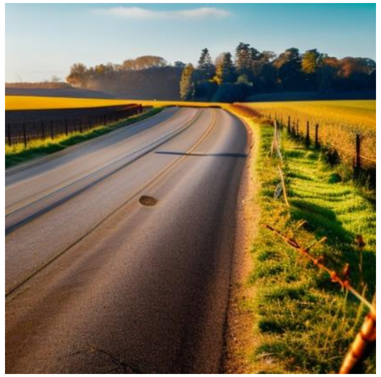




Class “rural landscape”

Positive prompt: Photograph of a rural landscape, high quality photography, Canon EOS R3

Negative prompt: digital art, drawing


Generating Data for semantic dependency analysis

Class “cityNoTrees” (less than normal presence of trees)

Positive prompt: Photograph a city, high quality photography, Canon EOS R3
Negative prompt: digital art, drawing, **trees**

Class “cityTrees” (more than normal presence of trees)

Positive prompt: Photograph a city, **trees,** high quality photography, Canon EOS R3
Negative prompt: digital art, drawing

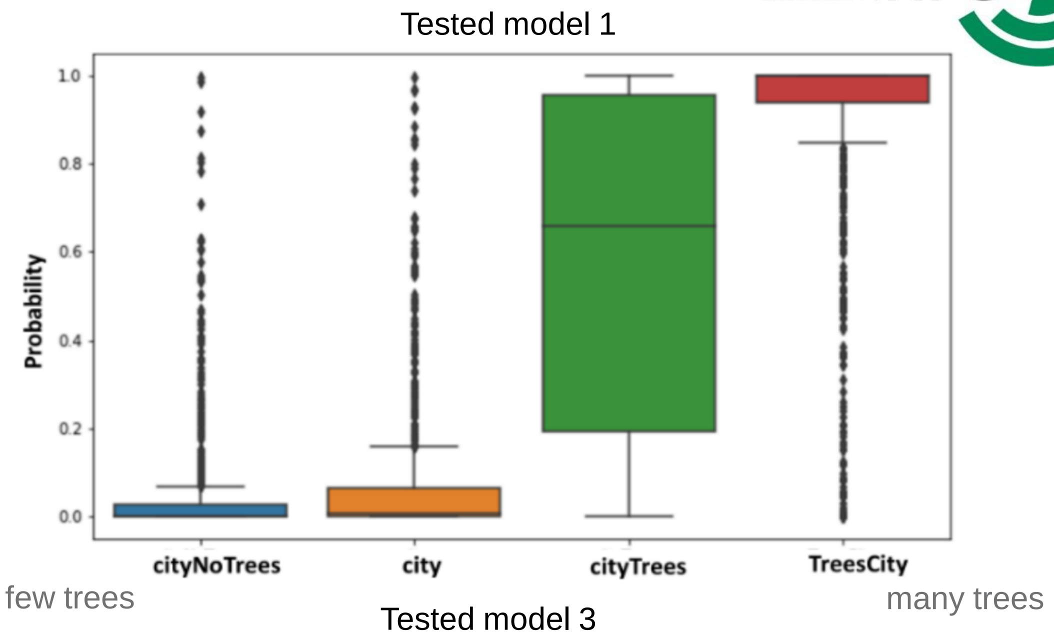
Generating Data for semantic dependency analysis

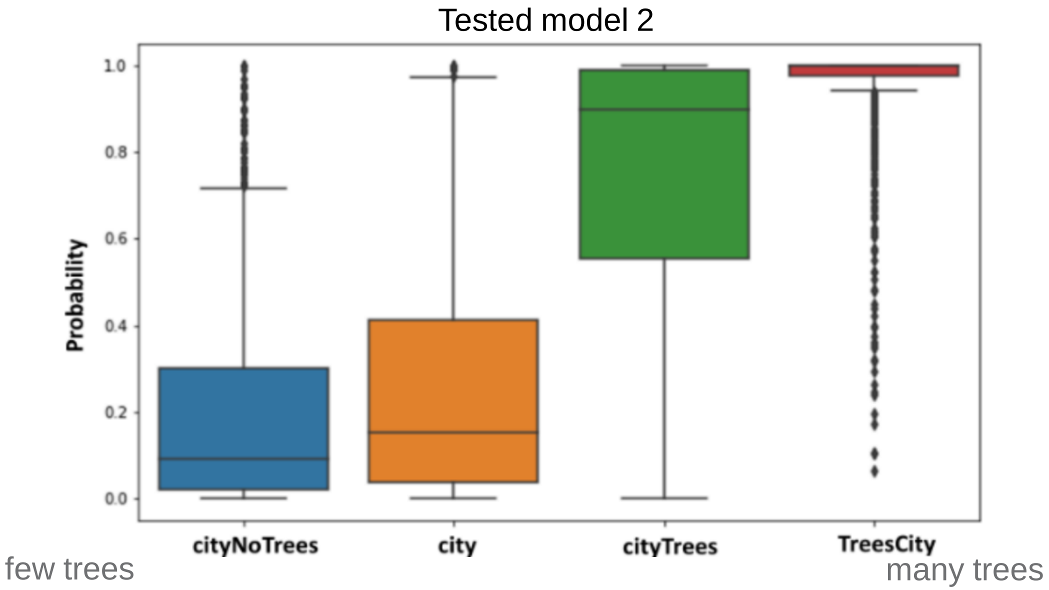
Class “TreesCity” (very high presence of trees)

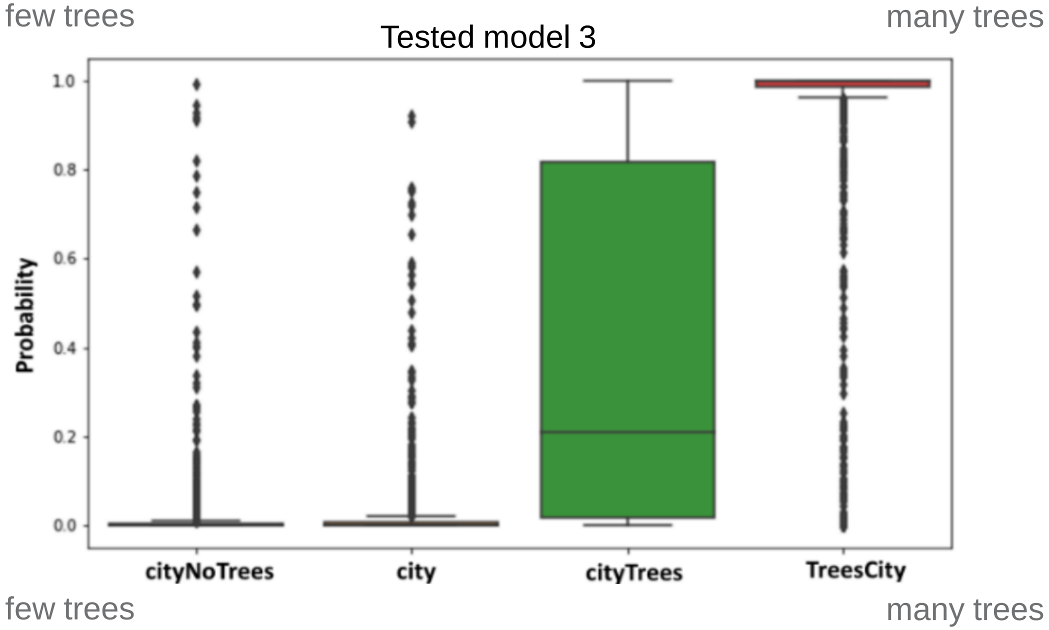
Positive prompt: Photograph **trees, city**, high quality photography, Canon EOS R3
Negative prompt: digital art, drawing

Experimental Results[1]

- SDA shows that the concept of “presence of trees” impact the classification in the expected way
- The plausible result validates the viability of the approach







[1]more results in the paper

Conclusion

- **Contributions**
  - We demonstrated a way to analyze model dependency on higher level concepts
  - We described two general ways for implementation (trough generators and detectors)
  - In experiments we demonstrate the feasibility in a sample implementation
- **Challenges**
  - Implementing generators and detectors with the desired behavior
- **Future work**
  - Exploring implementation of generators and detectors (e.g. 3D engines, diffusion models with image to image, etc.)