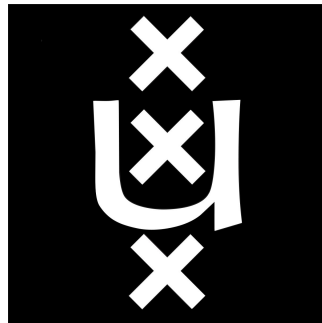


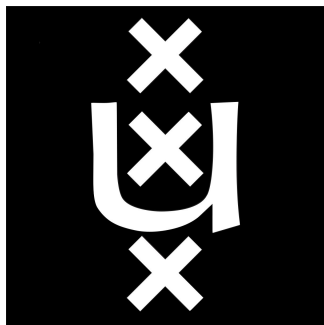
AI for Society

BSc AI 2020/21



Week 2: eXplainable AI

Giovanni Colavizza

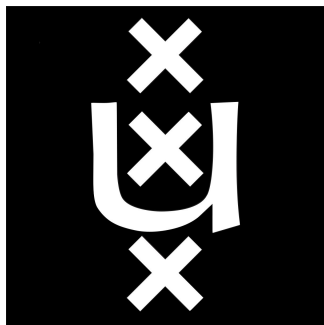


Topics

1. Bias in word embeddings
2. Defining XAI
3. LIME
4. Assignment

PART 1: Bias in word embeddings

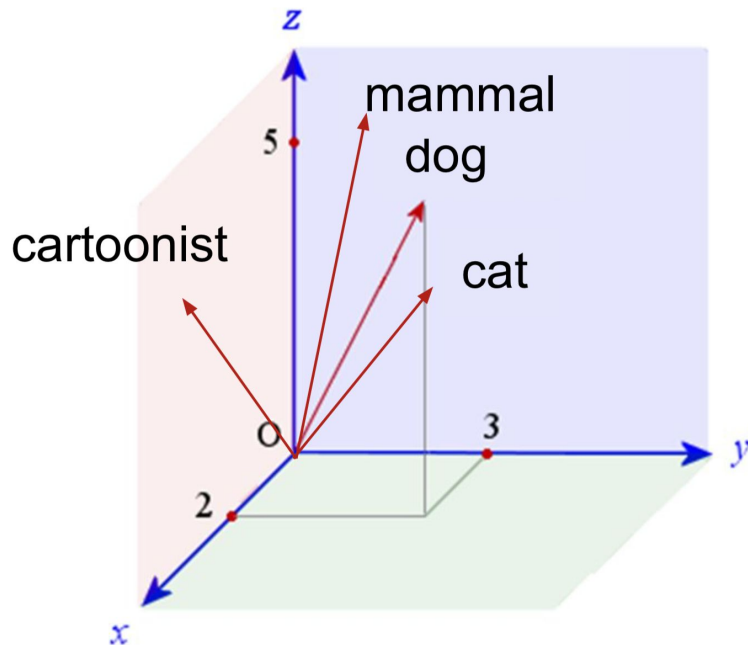
Giovanni Colavizza



Word embeddings

A numerical representation of linguistic units, usually words, as vectors in a high-dimensional space.






Word embeddings can be very *sparse*, e.g., using counts, or *dense*, e.g., using modern deep learning techniques (Word2vec, BERT).



Count-based word embeddings

Term-document
matrix: sparse.

Can be made
dense via
dimensionality
reduction (e.g.,
using SVD).

					
dog	3	10	0	0	0
Internet	4	0	1	0	7
cartoon	1	0	5	0	0
cartoonist	1	0	10	0	0

term

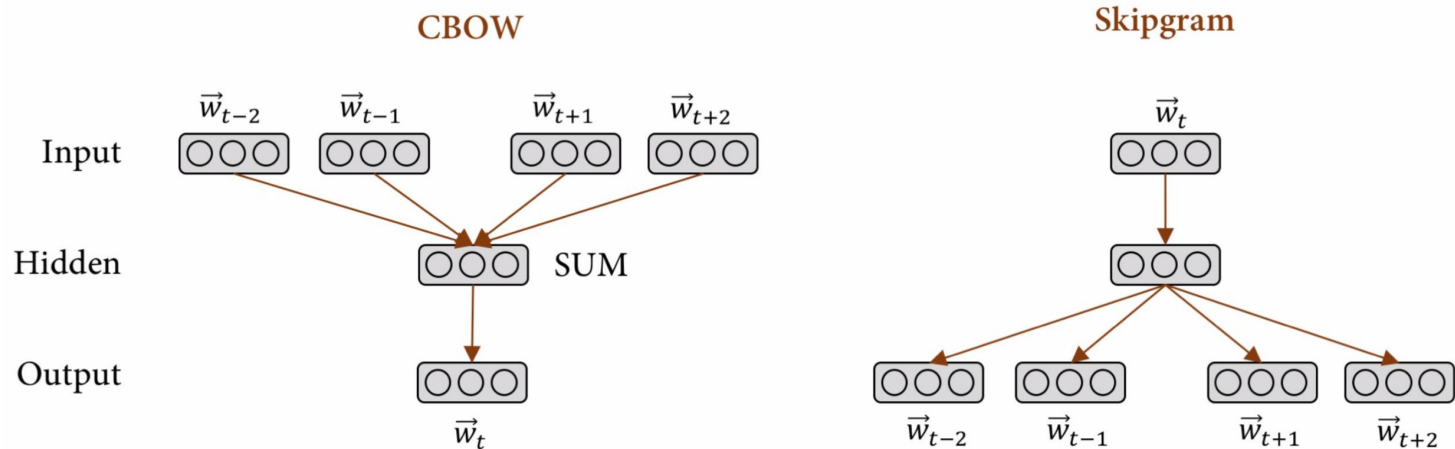
document

count of term occurrences
within the document

Prediction-based word embeddings

Word2Vec; Mikolov et al. 2013

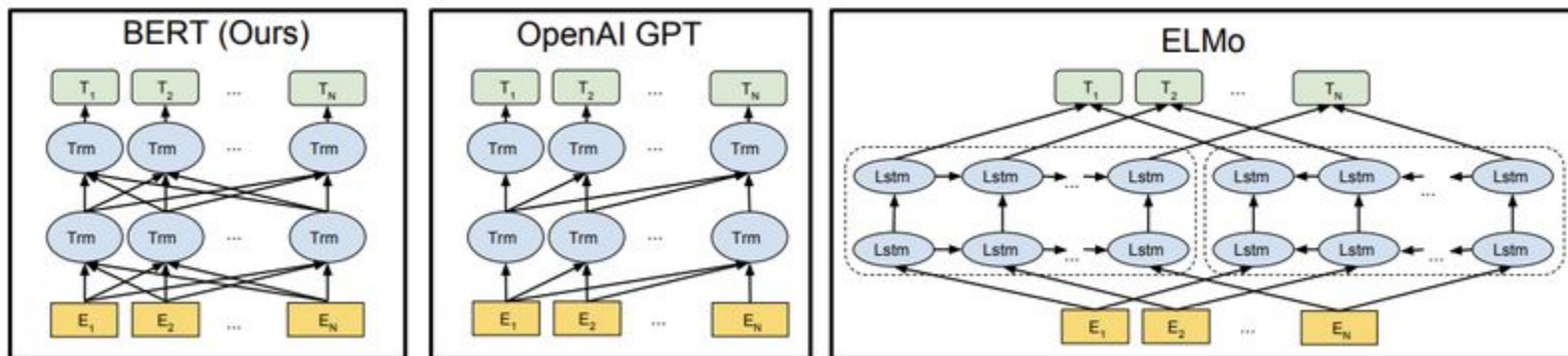
<https://arxiv.org/pdf/1301.3781.pdf>



Attention-based word embeddings

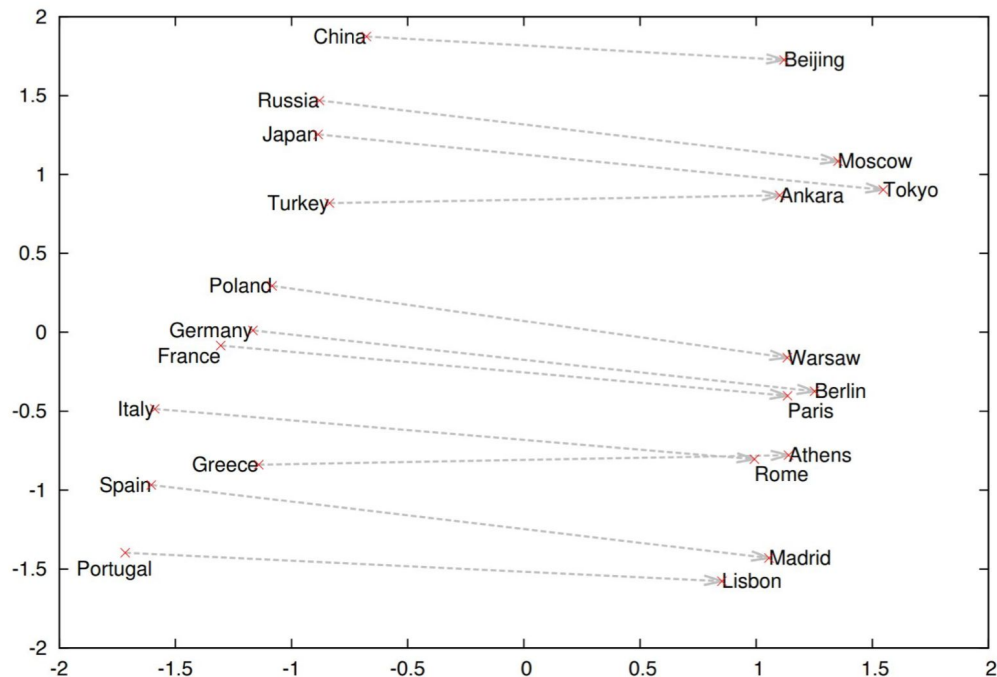
BERT; Devlin et al. 2019

<https://arxiv.org/pdf/1810.04805.pdf>

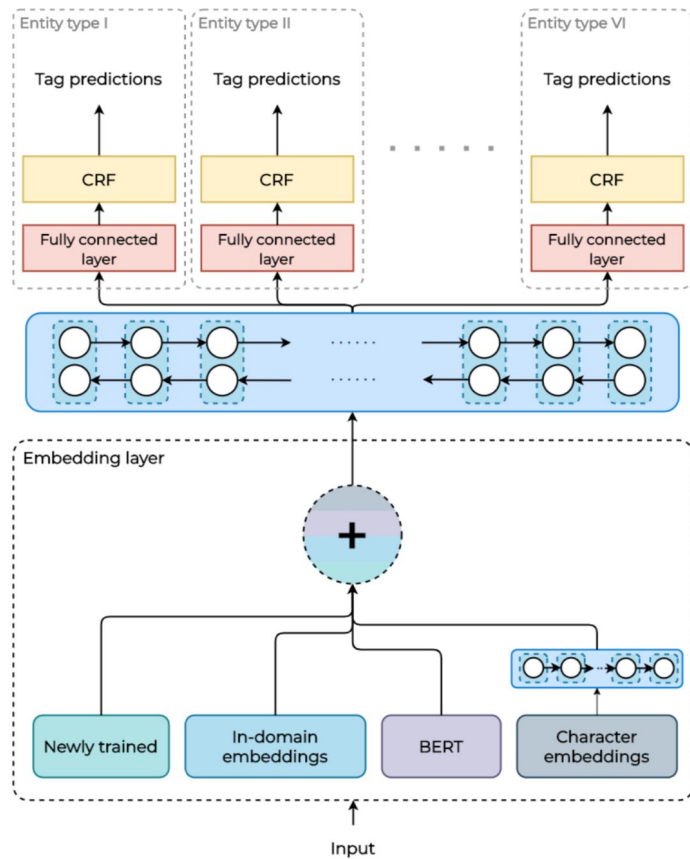


Word embedding analogies

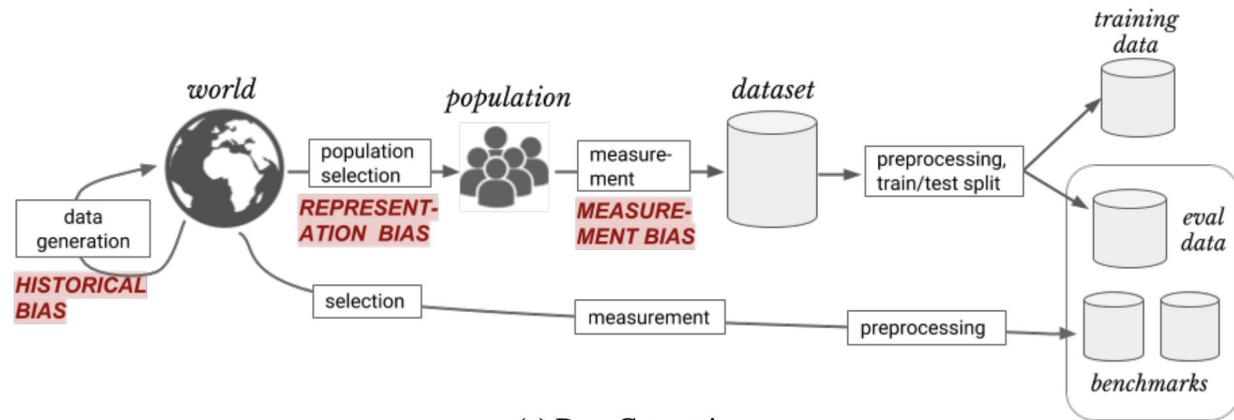
Rome - Italy +
Greece = Athens



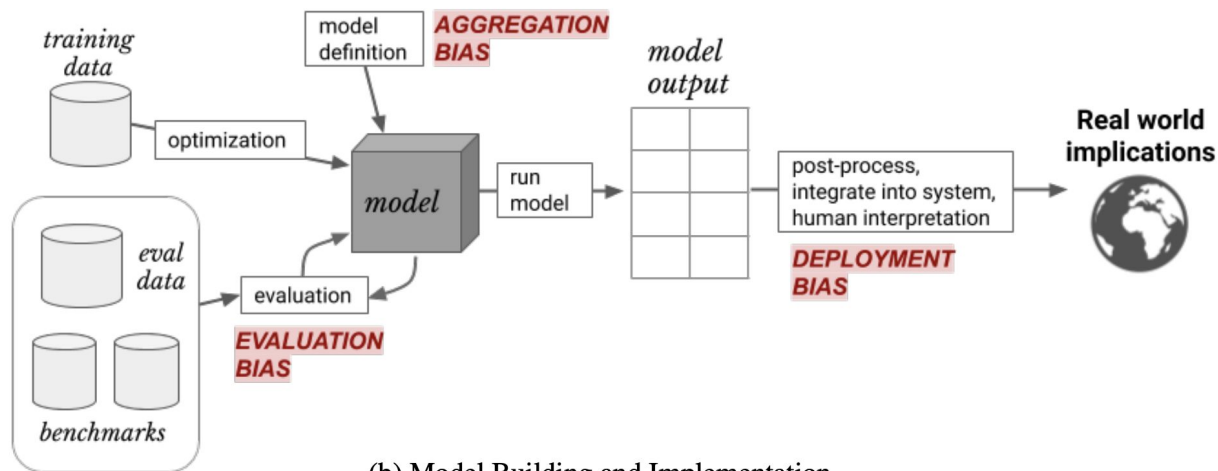
Word embedding as features



Bias in AI

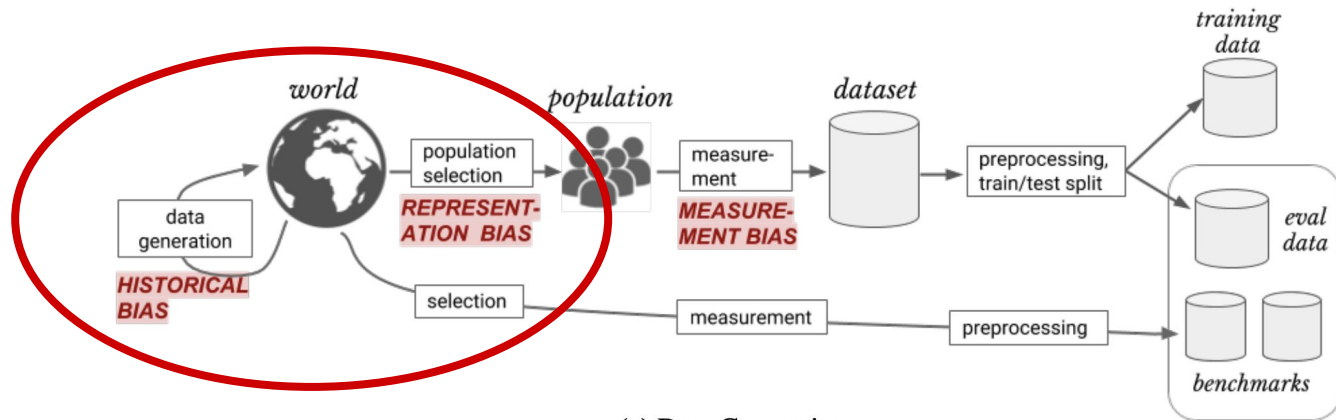


(a) Data Generation

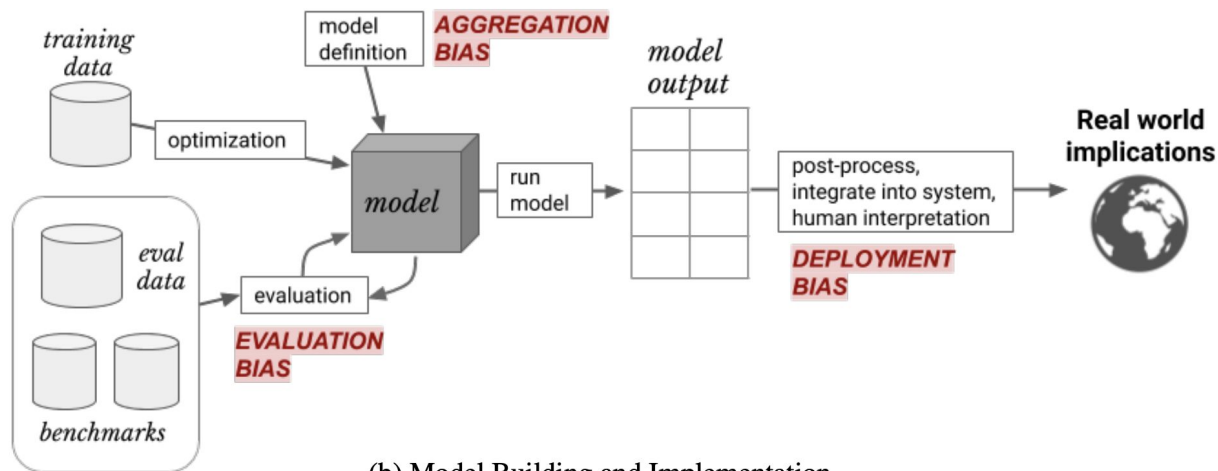


(b) Model Building and Implementation

Bias in AI

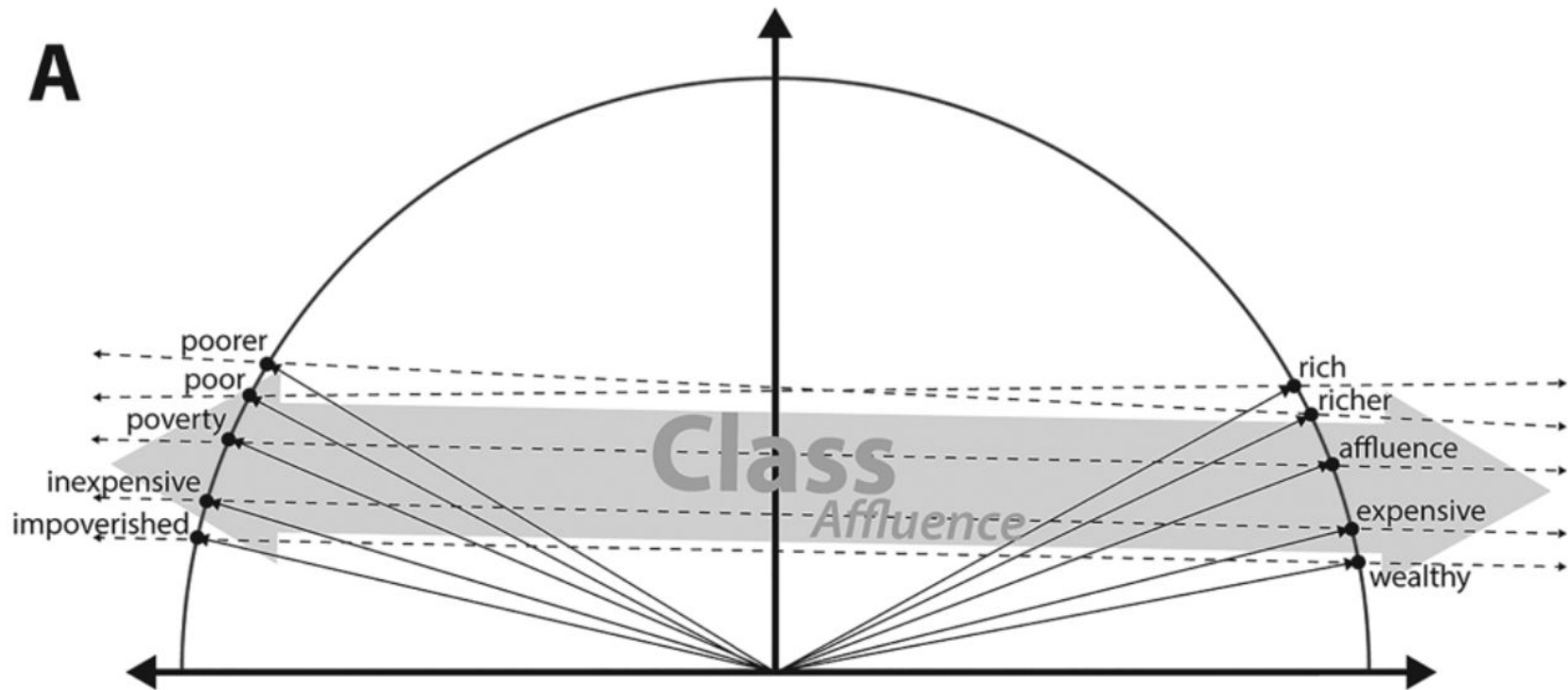


(a) Data Generation

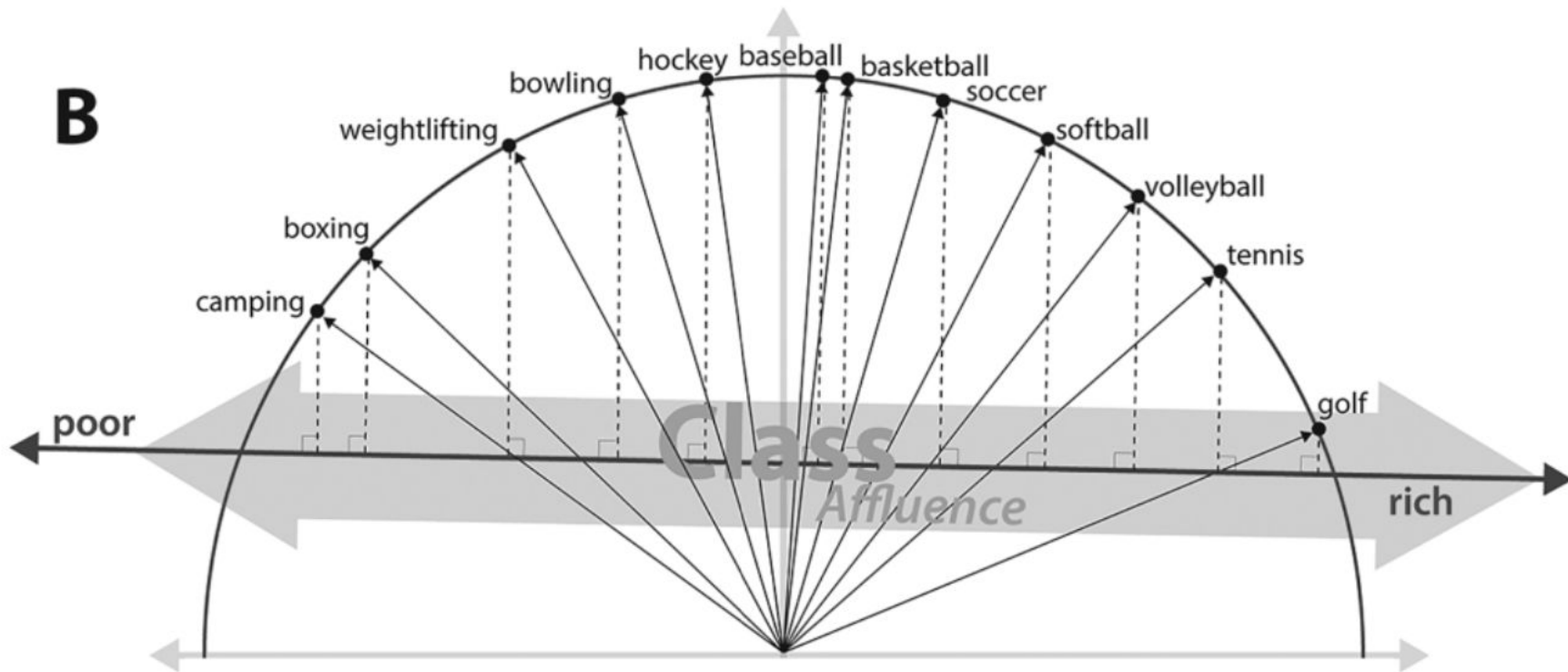


(b) Model Building and Implementation

Bias in word embeddings



Bias in word embeddings



Measuring (gender) bias in word embeddings

Intuition:

- Define a set of “definitional word pairs” that capture the gender dimension (e.g., he/she, man/woman, etc.)
- Measure bias by how differently a word w projects onto word pairs. For example, consider “politician”. First, calculate $x_{he} = \cos(\text{“politician”, “he”})$ and $x_{she} = \cos(\text{“politician”, “she”})$. Then, use $x_{he} - x_{she}$ to get a measure of bias towards the masculine gender.
- Note: *cosine* is a common measure among vectors where magnitude is not important, but we only care about the direction (angle).

Measuring (gender) bias in word embeddings

Identify the *gender subspace*:

- Consider the pairwise differences among the set of “definitional word pairs” that capture the gender dimension (he - she, man - woman, etc.)
- Apply dimensionality reduction on them, and find the gender subspace. The authors use PCA and find that the first PC is sufficient for this.
- Use the *cosine* between any word and this gender subspace to quantify its bias. This bias can be averaged over a set of words.
- If you take masculine - feminine, a positive cosine might be indicative of bias towards the masculine gender, vice versa for a negative one.

Dealing with (gender) bias in word embeddings

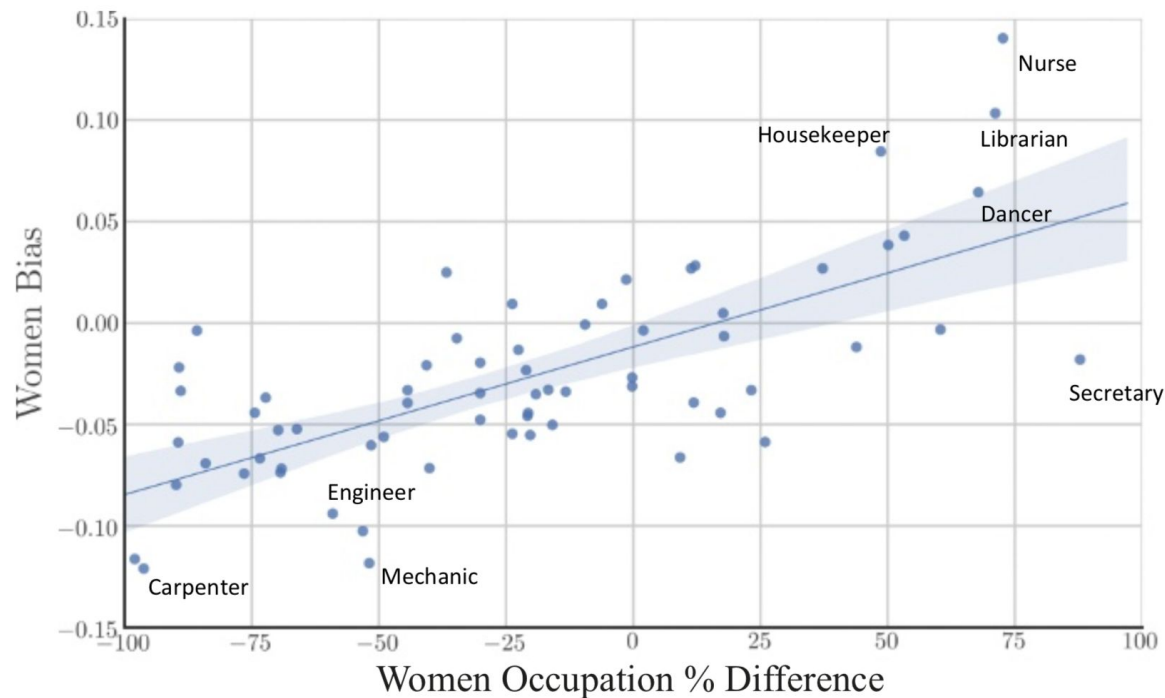
The authors propose two options:

- Neutralize and equalize (**hard de-biasing**): enforces that any gender neutral word is set to zero onto the gender subspace.
- Soften (**soft de-biasing**): ensures that neutral words are equidistant from equality sets. For example, it ensures that {brother, sister} and {husband, wife} are both equidistant from {babysitting}, although probably the latter set will still be closer than the former.

Measuring (gender) bias in word embeddings

Let's try this out. See the course repository, under assignment 2.

Bias in word embeddings



Approaches to bias in word embeddings

Three families of approaches:

1. Work on data (e.g., by filtering the training corpus)
2. Work on the algorithm (e.g., by adjusting the loss or factoring-in bias mitigation via a constrained optimization objective)
3. Work on the outcome, or post-hoc methods (e.g., de-bias word embeddings by transforming them in some way)

Sometimes, **1+2 are called model-based and 3 post-hoc approaches**. The method we discussed so far is a *post-hoc approach by example*.

Limitations and more work

Gonen and Goldberg. 2019. *Lipstik on a pig*, <https://arxiv.org/abs/1903.03862> (critique of Bolukbasi et al.)

Calyscan et al. 2016. *Semantics derived*, <https://arxiv.org/abs/1608.07187> (introduces the Word Embedding Association Test)

Kurita et al. 2019. *Measuring bias*, <https://arxiv.org/abs/1906.07337> (recent work on BERT)

Sun et al. 2019 *Literature review*, <https://www.aclweb.org/anthology/P19-1159>

Bias in word embeddings and XAI

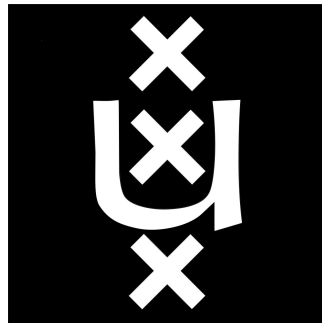
Detecting, quantifying, and dealing with bias in word embeddings brings us to **eXplainable AI**. In general, we want to:

- Understand if something is happening (is there gender bias?)
- Quantify and qualify the effect (how much bias? Where?)
- Apply countermeasures and verify (is there still bias after we applied a countermeasure? Is my classifier using word embeddings as features, thus relying on biased signals to make decisions?)

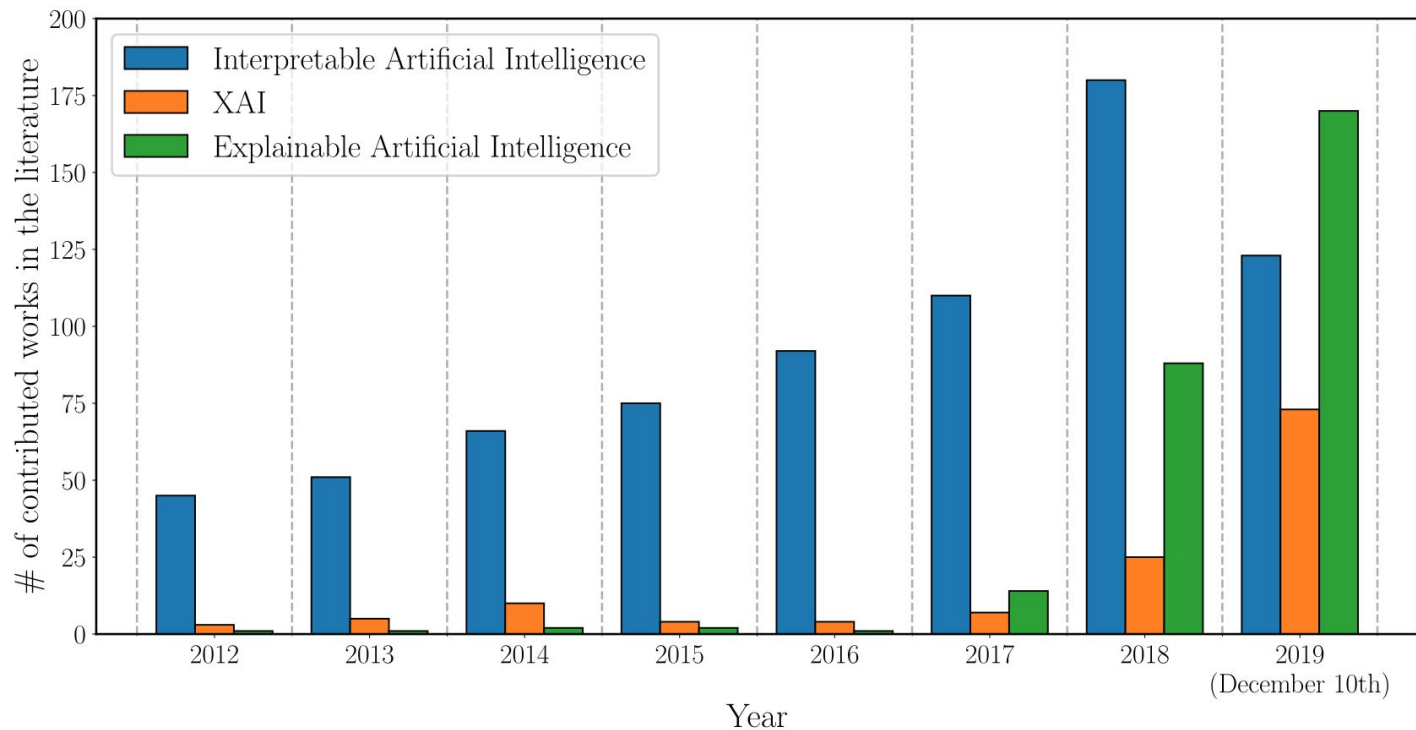
Q&A

PART 2: Defining XAI

Giovanni Colavizza



Another recent trend..



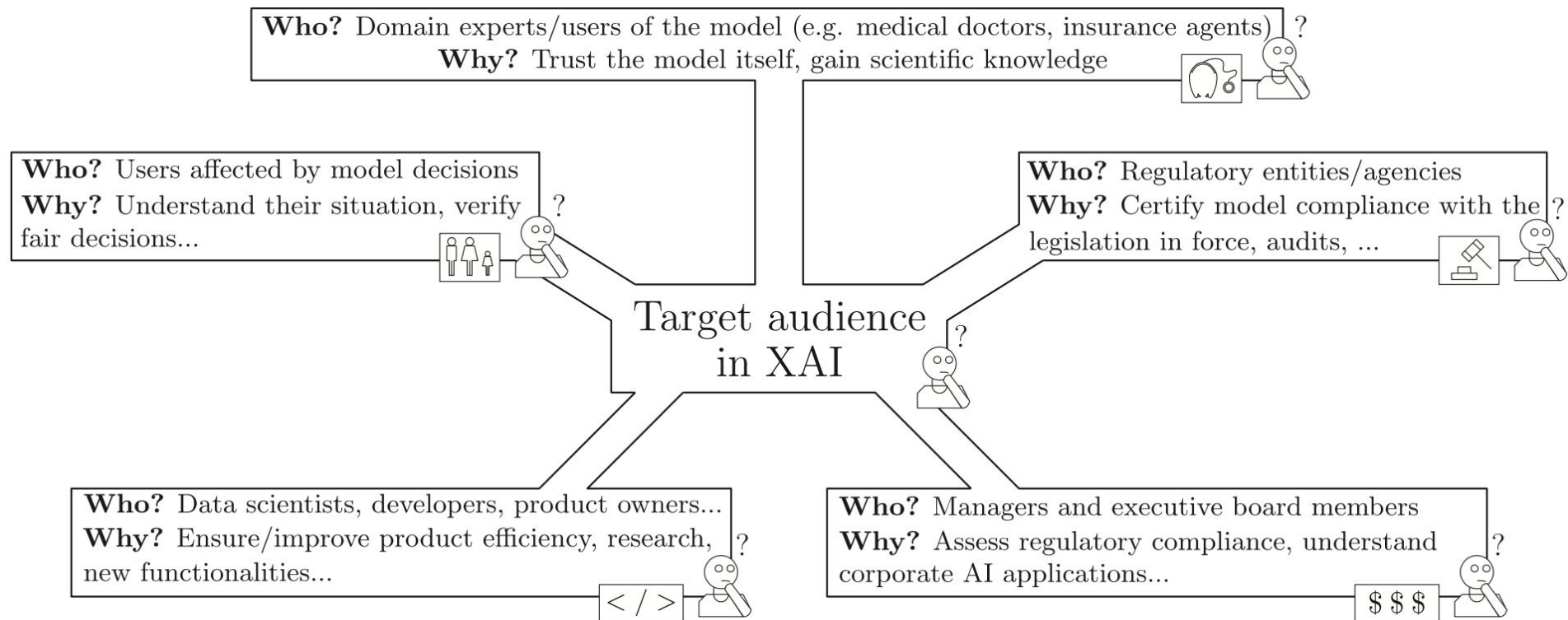
Towards a definition

- Not always clear what people refer to when discussing explainable or interpretable AI. The two are often interchanged.
- We will distinguish as follows:
 - **Interpretations** focus on **model transparency**, and ask the question: “how does a model work?”
 - **Explanations** focus on **post-hoc reasoning**, and ask the question: “what can a model tell me?”
- Note: the literature is ambiguous, but hopefully these definitions are not (although they overlap and are clearly related). You will need to understand what is what on a case by case basis.

Towards a definition

- Following upon our definitions, we can say that an **interpretable model** is one that allows a human to, locally or globally, grasp the mechanism used to come up with a decision. **Being interpretable is a property of a model**. An example is a simple linear regression model (without too many parameters and transformations), the logic of which we can fully grasp.
- Instead, **explaining a model** attempts to reconstruct the logic (e.g., steps) which a model uses to come up with a decision. **Explaining is an action that we take, not a property of a model**. An example is a deep neural network defining a complex decision boundary, which we attempt to explain by perturbing the inputs while observing the outputs.

Explanations, but for whom?



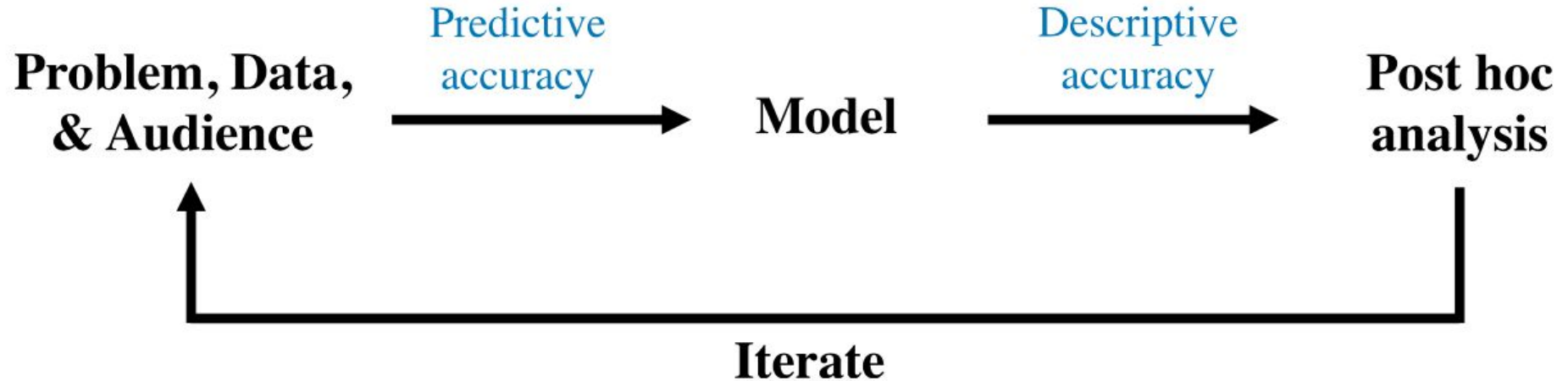
Predictive, descriptive, relevant

Which explanations/interpretations are good? Those that are

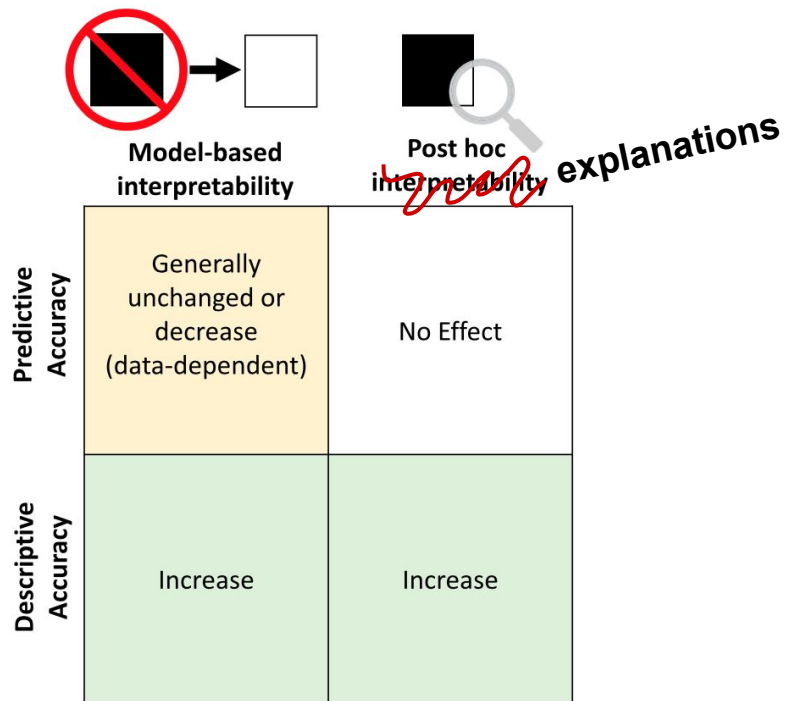
- **Predictive**, in that the model which we explain/interpret works well at the task at hand. This is the normal evaluation of ML models (e.g., accuracy).
- **Descriptive**, in that they accurately map the model's decisions to humans.
- **Relevant**, in that they are useful to a human in a certain setting/domain.

A visual interpretation by example of a deep neural network boundary could be highly predictive if the network works well for the task at hand, but poorly descriptive if the explanation does not map well to decisions and hence does not convey insights to the human observer on how they were taken.

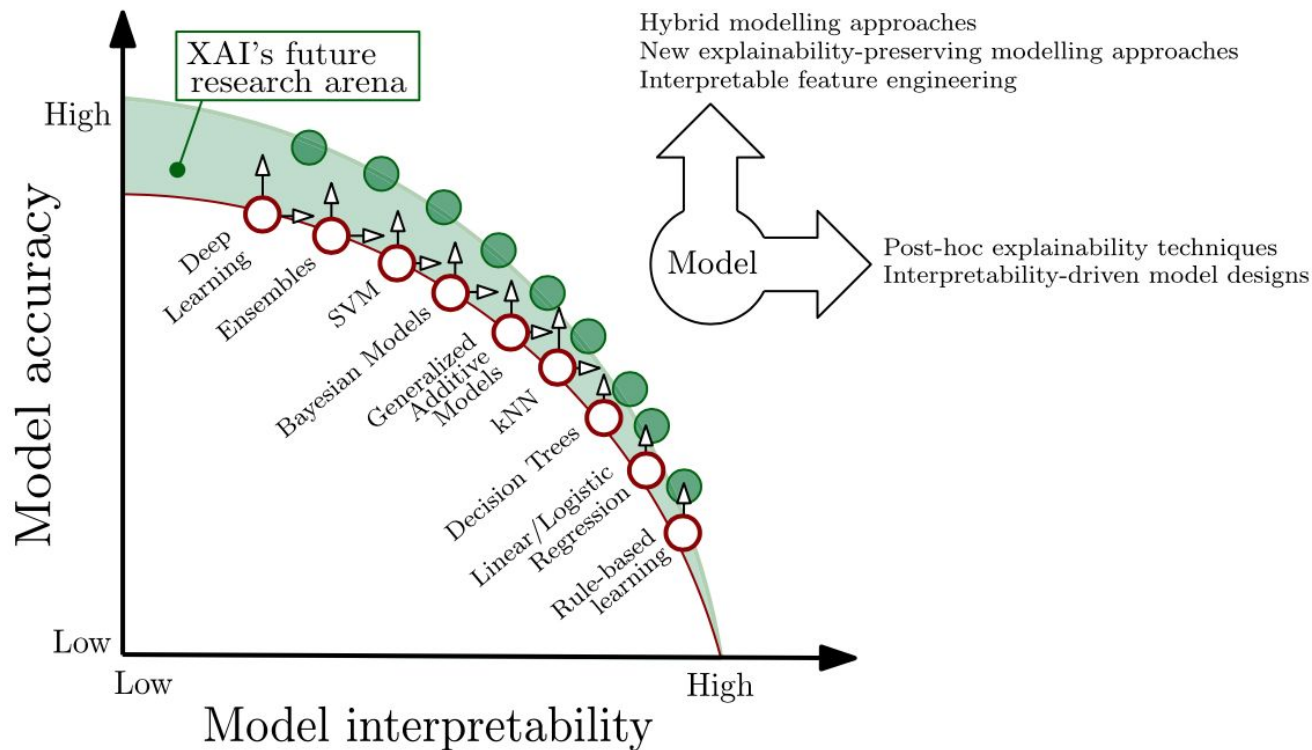
Predictive, descriptive, relevant



Performance vs Interpretability tradeoff



Performance vs Interpretability tradeoff



Traits of an explanation/interpretation

Explanations/Interpretations are Human-AI interactions, therefore they also have further “social” traits we should keep in mind:

- **Confidence:** grows when the rationale of a decision is close to the thought processes of the user.
- **Trust:** grows when decisions do not require validation to be acted upon.
- **Safety:** an automated decision-making system which is consistent and reliable, sends warnings when uncertain and anyway does show degree of confidence, is robust to perturbations and outliers, can fail well, etc.
- **Ethics:** an automated decision-making system which does not violate a certain well-defined code of principles.

Traits of an explanation/interpretation

Explanations/Interpretations are not only causal, but also *contextual*, in that they need to serve a human in a certain situation. Insights from the social sciences, psychology and philosophy highlight that explanations should also:

- **Be contrastive**
- **Be selective**
- **Provide causes, more than probabilities**
- **Consider the social context**

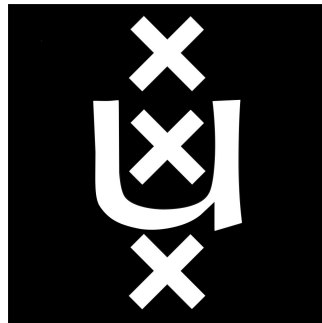
The jury is still out there

- Explainable vs Interpretable AI
- Performance vs Explainability/Interpretability tradeoff
- Explanations *in context*
- Regulatory frameworks (or lack thereof)

Q&A

PART 3: LIME

Giovanni Colavizza



Local Interpretable Model-agnostic Explanations

Requirements:

- “Explanations must be **interpretable**, i.e., provide qualitative understanding between the input variables and the response”
- “Explanations must be **locally faithful**, i.e., correspond to how the model behaves in the vicinity of the instance being predicted”
- “The explainer should be **model-agnostic**”
- “Explanations should provide a **global perspective**” on the behaviour of the model

Local Interpretable Model-agnostic Explanations

LIME, an algorithm that can explain the predictions of any classifier or regressor in a faithful way, **by approximating it locally with an interpretable model.**

SP-LIME, a method that **selects a set of representative instances** with explanations to address the “trusting the model” problem.

“The overall goal of LIME is to identify an interpretable model over the interpretable representation that is locally faithful to the classifier.”

LIME's cost function for a local explanation

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

Explanation for datapoint x

Class of interpretable models. E.g., a linear model.

Measure of how well g approximates f at the locality defined by π_x

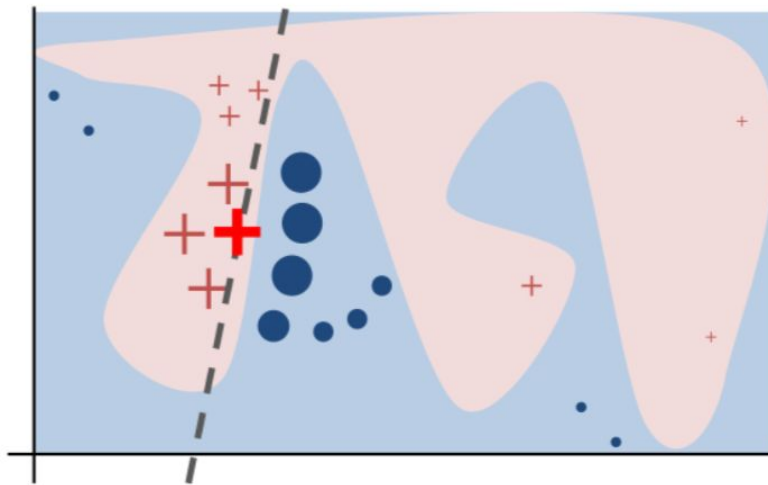
Fitted model to be explained

Measure of proximity of a datapoint z to x

Complexity of g . For example, number of parameters of a linear model

LIME's intuition

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

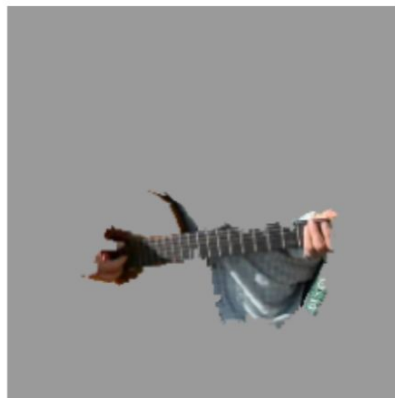


LIME in action

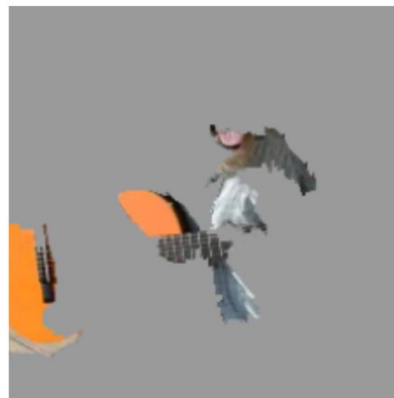
$$\xi(x) = \operatorname{argmin}_{g \in G} \mathcal{L}(f, g, \pi_x) + \Omega(g)$$



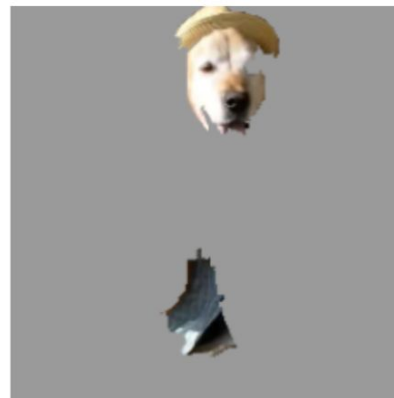
(a) Original Image



(b) Explaining *Electric guitar*



(c) Explaining *Acoustic guitar*



(d) Explaining *Labrador*

LIME in action

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

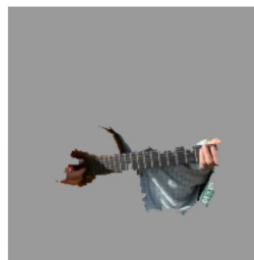
Where:

- G is the set of linear models
- L is a weighted square loss
- π_x is an exponential kernel defined on an L2 distance (Euclidean)
- Ω is the number of super-pixels (which are the features of the linear model)

An explanation is given using super-pixels which locally approximate f well as modelled by g .



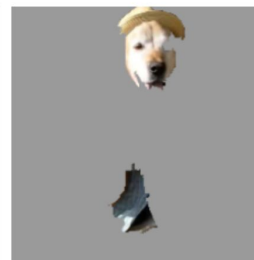
(a) Original Image



(b) Explaining *Electric guitar*



(c) Explaining *Acoustic guitar*



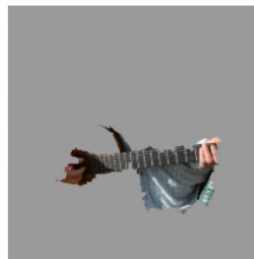
(d) Explaining *Labrador*

From local to global

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



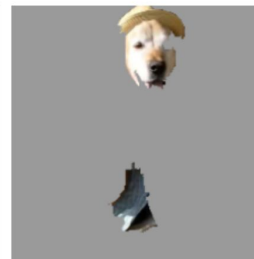
(a) Original Image



(b) Explaining *Electric guitar*



(c) Explaining *Acoustic guitar*



(d) Explaining *Labrador*

SP-LIME is a method to select which data points x to use in order to convey to the human a trustworthy representation of how the model behaves globally.

You can read the details in the original paper, including a validation with human annotators.

LIME

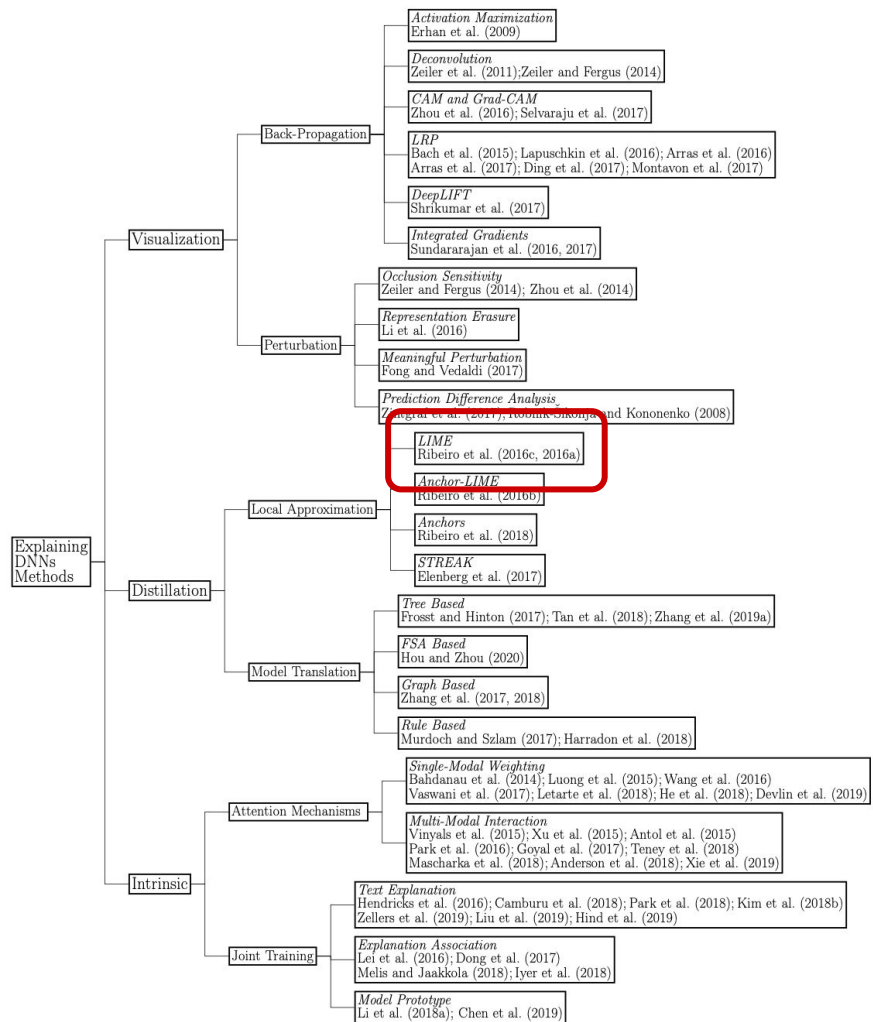


Figure 4: Methods for explaining DNNs.

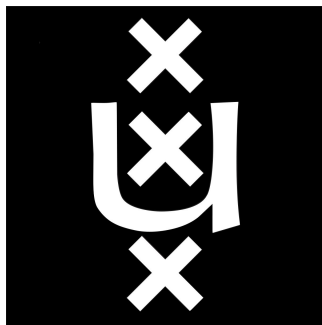
There is more out there

- Awesome list of libraries to explain black-box models:
<https://github.com/EthicalML/awesome-production-machine-learning#explaining-black-box-models-and-datasets>
- LIME implementation in Python (which you will use for the assignment):
<https://github.com/marcotcr/lime>
- SHAP unifies LIME and many more methods, formally and in this library:
<https://github.com/slundberg/shap>
- AIX360: <https://github.com/Trusted-AI/AIX360>
- Language Interpretability Tool (from UvA): <https://github.com/pair-code/lit>

Q&A

PART 4: Assignment

Giovanni Colavizza



Civility in online communication

Before you...



THINK!

Set-up

You will work into groups. Motivated requests to change group can be made.

Let's check the course repository for more info (assignment 2):

https://github.com/Giovanni1085/UvA_AlforSociety_2021

Note: we will assume you can clone a GitHub repository, set-up a working Python environment (ideally virtual, e.g., via Conda) and work with Jupyter notebooks. If you need some pointers/help, we have included a guide to setting up your working environment in the repo.

Q&A