Human-in-the-loop approaches to XAI

AIRA seminar 20.10.2022

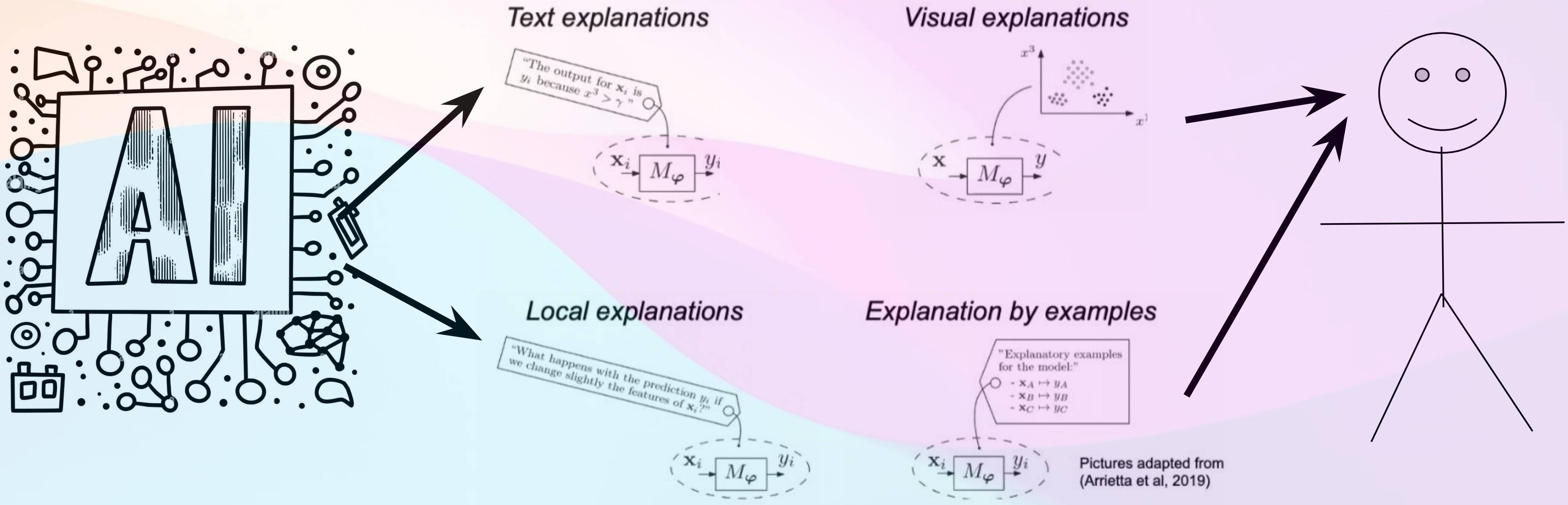
JAGIELLONIAN UNIVERSITY IN KRAKÓW

GEIST Research Group

We are GEIST. We dream big and work hard.

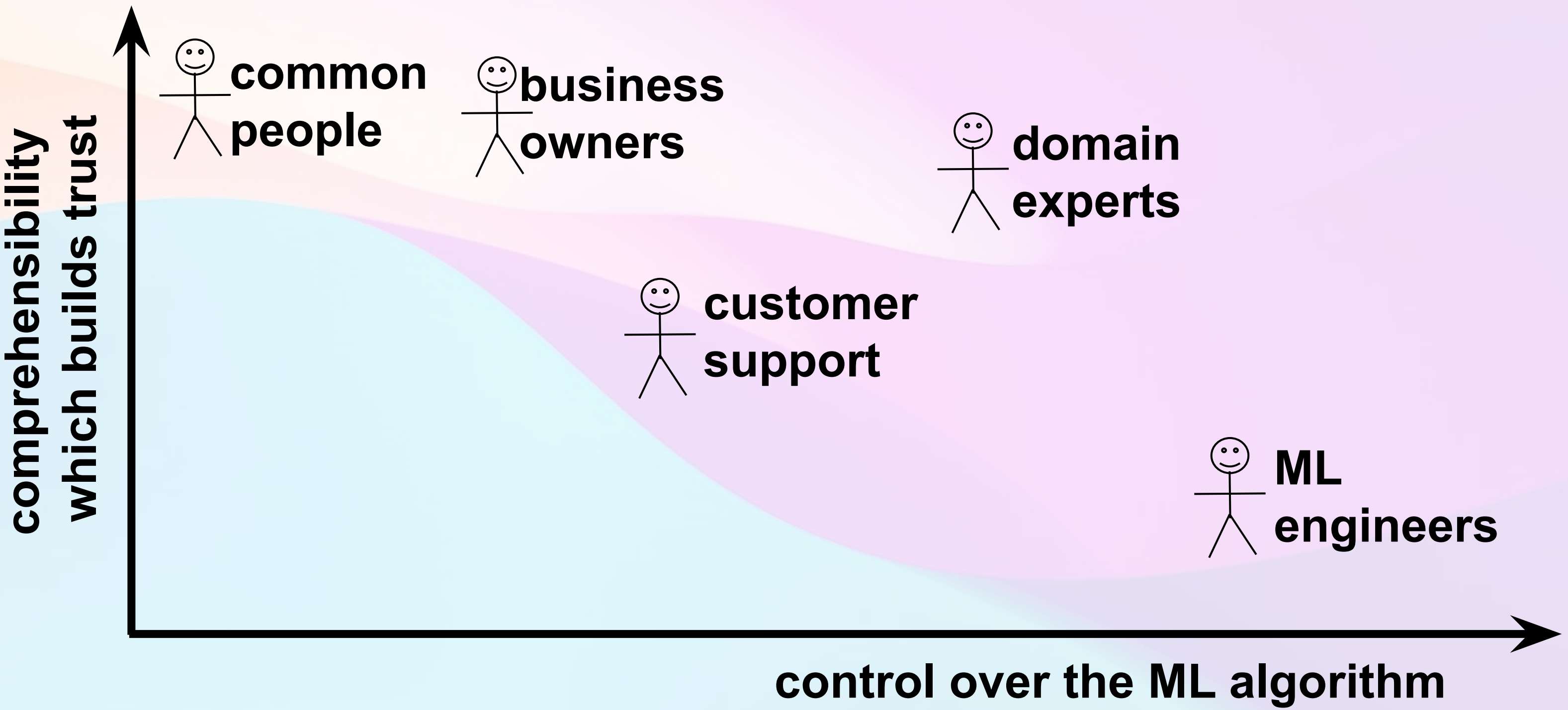
1. Explainable Artificial Intelligence
2. Human-in-the-Loop approach
   a. Objective data & metadata
   b. Interactive clustering
3. Intelligible eXplainable AI framework
   a. Metrics for explainers
   b. Time Series extension

EXPLAINABLE AI



Pictures adapted from (Arrietta et al, 2019)

XAI approach tailored for specific audience



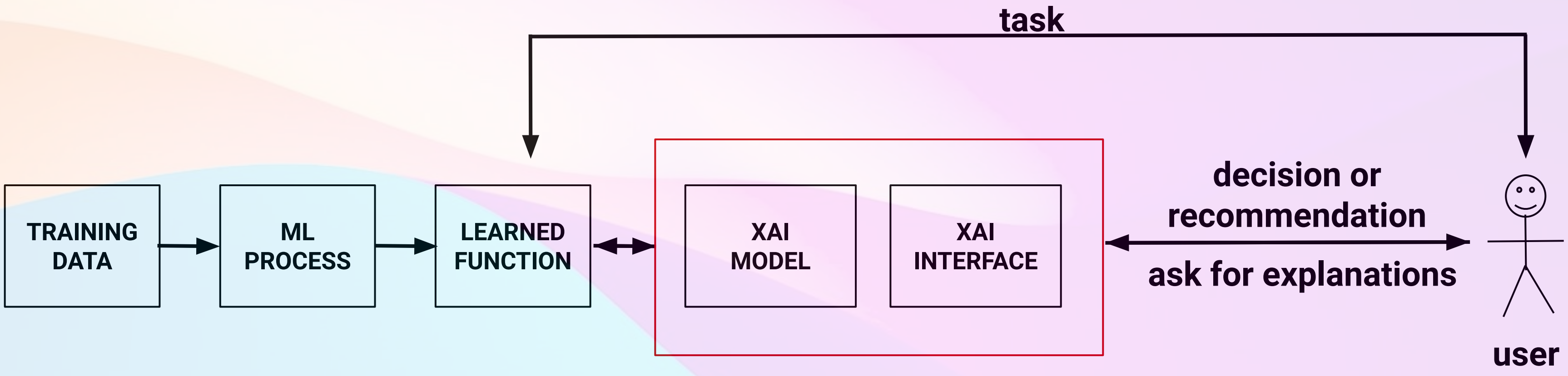
How to build trust in EXPLAINABLE AI?

Human-in-the-Loop approach

How to build trust?

Du, M., Liu, N., & Hu, X. (2019). Techniques for interpretable ML. Communications of the ACM, 63(1), 68–77.

Human-in-the-Loop



Understand why & why not

Know when to trust AI

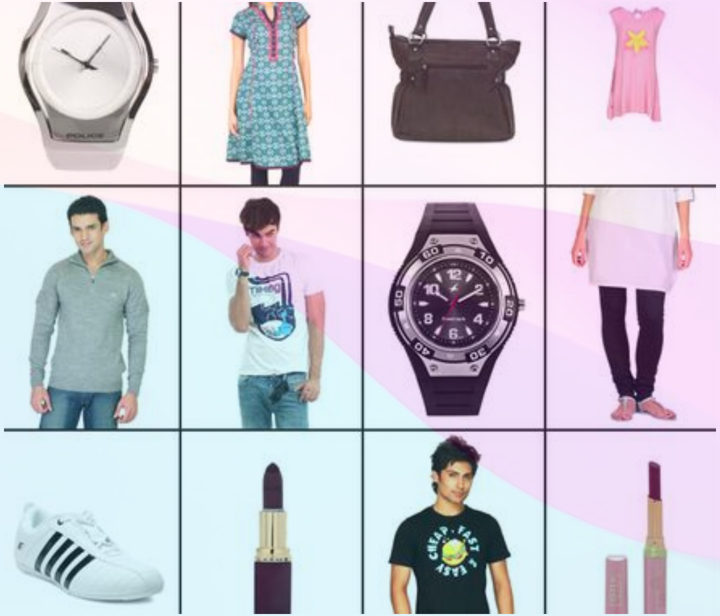
Is model fair?

Du, M., Liu, N., & Hu, X. (2019). Techniques for interpretable ML. Communications of the ACM, [63(1), 68–77](#).

eCommerce example

eCommerce example

objective data

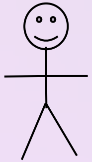


metadata

David Beckham
Signature Men Deos
2010

Fila Men's Round Neck
Navy Blue T-shirt
Autumn 2012

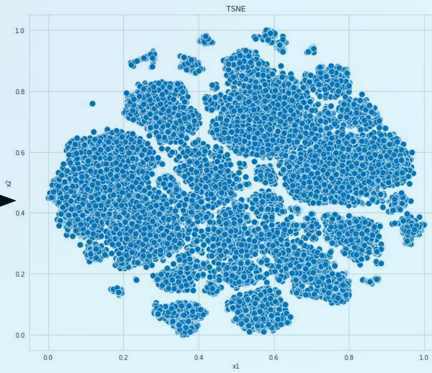
CASIO EDIFICE Men
Black Dial Chronograph
Watch ED60

	objective data	**metadata** 
Description	data derived directly from measurements (empiricists' approach)	"data about data", descriptions, labels
Specification	closer to the process under study, it is less subject to interpretation, needs explanation	more prone to errors and interpretation, laborious to create them, textual, allows for the formulation of an explanation
Usage	as independent variables in ML models: classification, regression, clustering, …	as explanations of ML models, target variables in ML models, sometimes independent variables
Examples	industrial sensors data, images of classes (eg. products, animals), video, medical data (eg. EEG, CT)	description of objects under study, how the measurement was carried out / data collected

Objective data

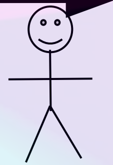
Interactive clustering: select # of clusters

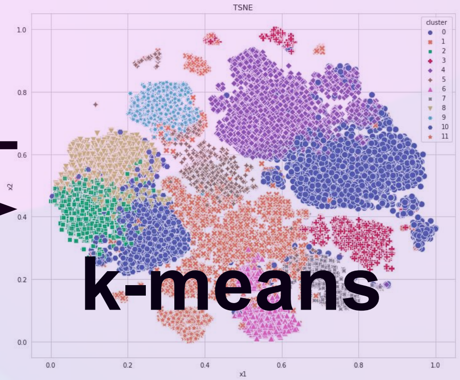
MobileNet → *[0.123, 0.456, ...]* → $\mathbb{R}^{20k}$ $\xrightarrow{\text{SVD}}$ $\mathbb{R}^{3k}$ $\xrightarrow{\text{TSNE}}$



+

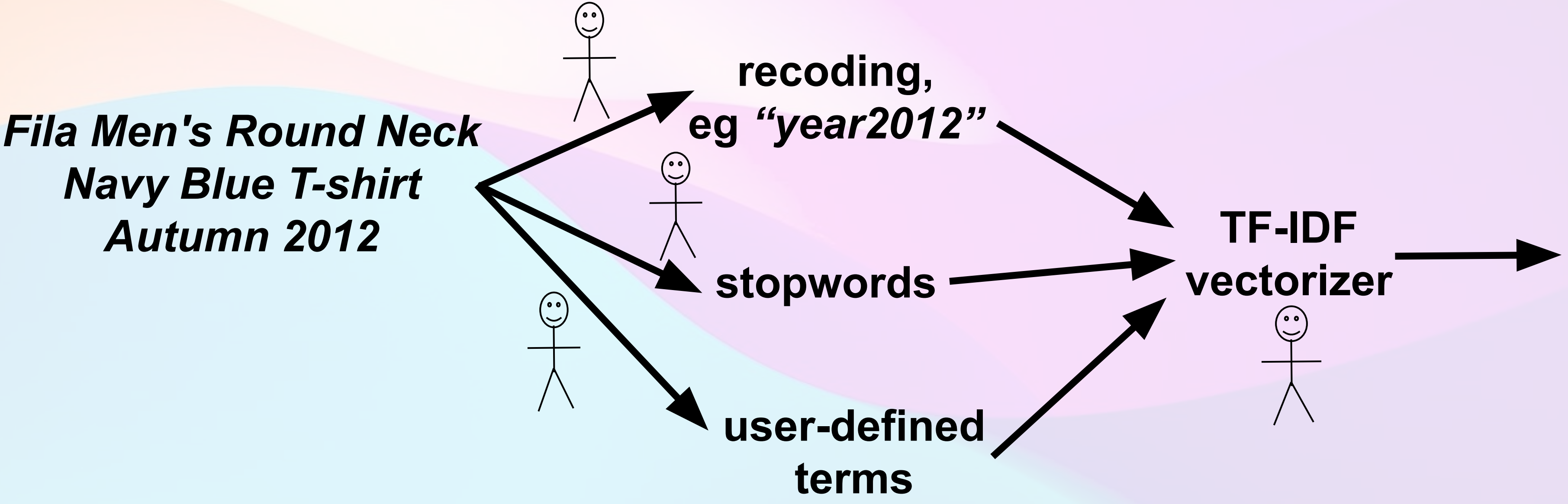
Silhouette score





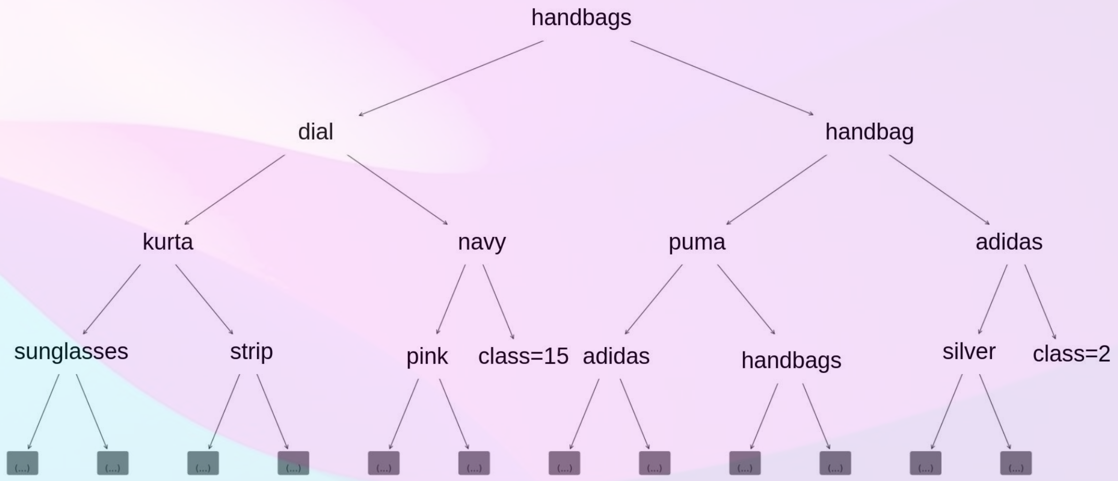
k-means

Metadata
Inspiration: Natural Language Processing



Metadata
Inspiration: Natural Language Processing

TF-IDF vectorizer



decision tree classifier

Metadata

Inspiration: Natural Language Processing

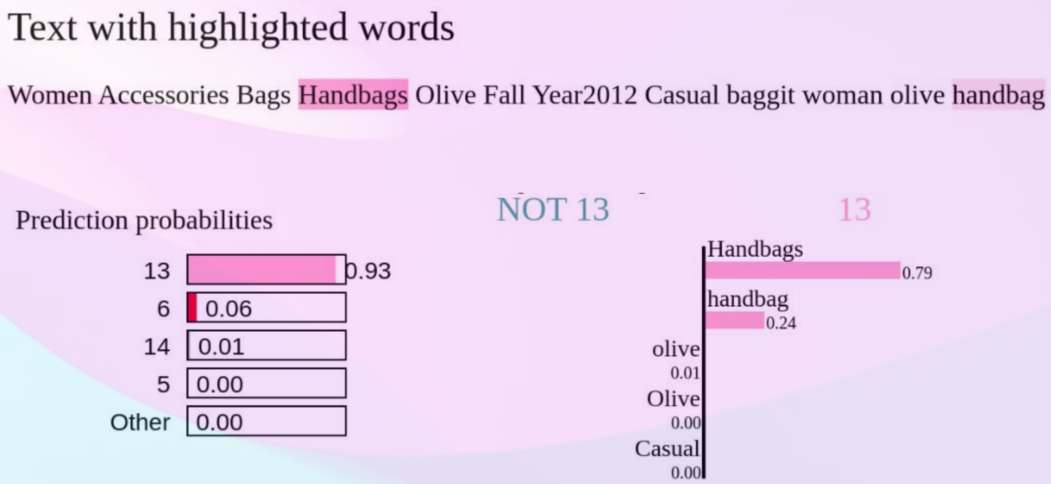
TF-IDF vectorizer

word cloud

Metadata

Inspiration: Natural Language Processing

TF-IDF vectorizer



LIME explanation

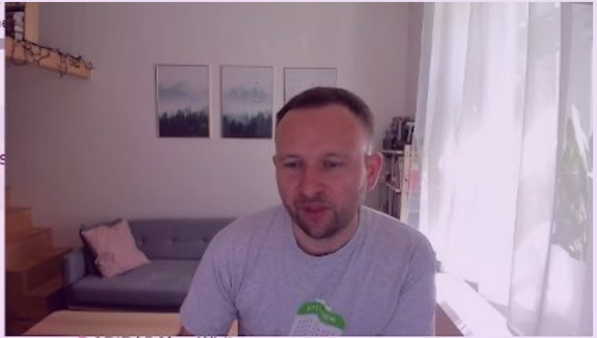
rel	Topwear	Tshirts	Black	Summer	2011	Casual	Billy Idol Men's Faceless Blac...	billy idol me

5000 rows × 16 columns Open in new tab

In 33

```
_pipeline.visualise_clustering(column_year = 'year', column_txt = 'productDisplayName', random_s
```

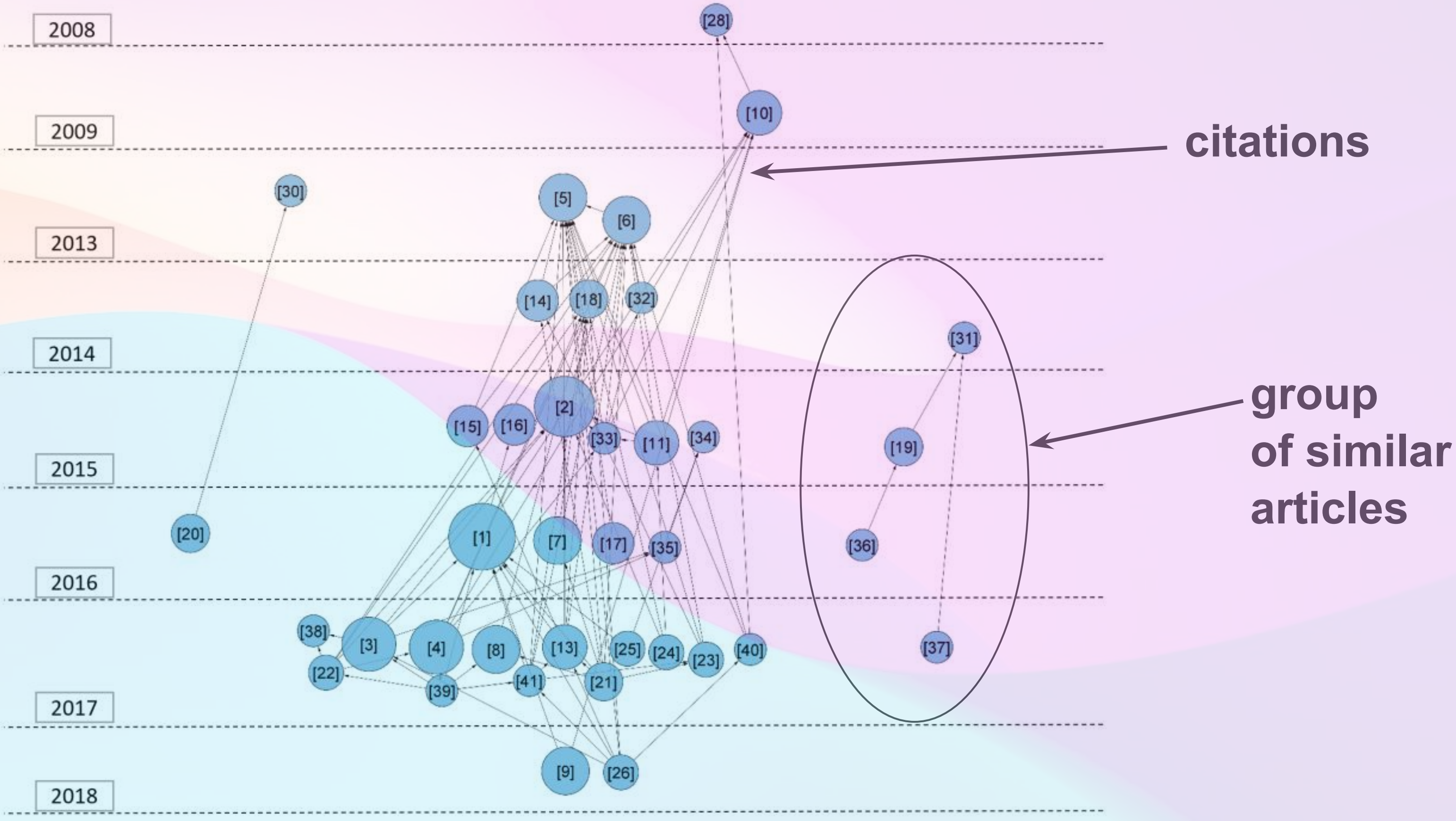
subsample, size factor



IJCAI-ECAI 2022 Workshop on semantic techniques for narrative-based understanding

XAI Survey example

Field's Evolution Graph



XAI Survey - 2 types of data

objective data

	article1	article2	article3	...
article1	-			
article2	1	-	1	
article3		1	-	
...				

metadata

HITL Summary

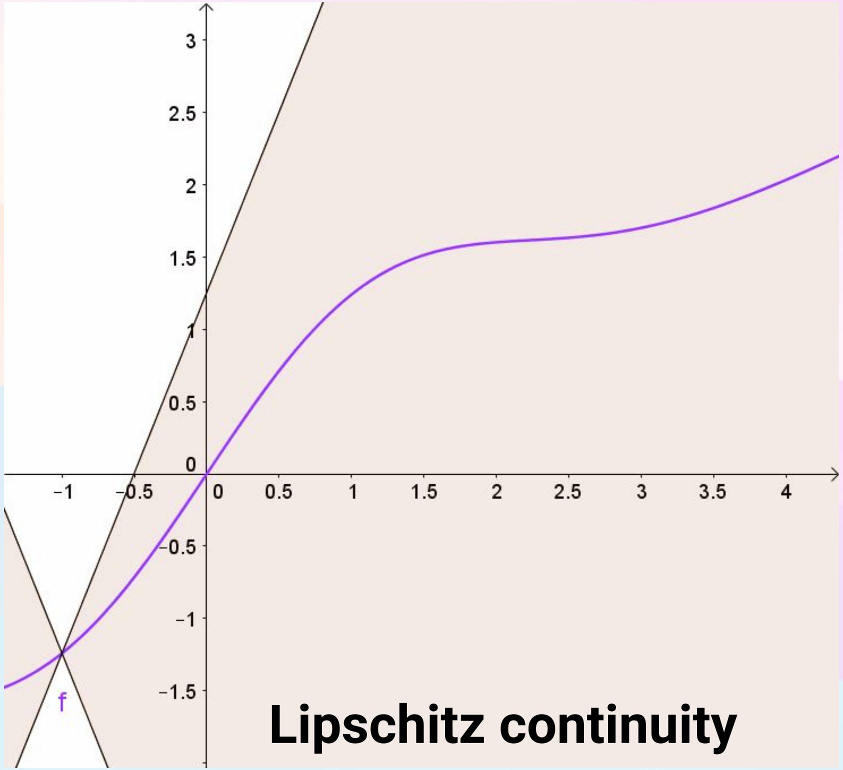
XAI is an intermediary layer between the ML algorithm and the human. It should be tailored for specific audience.

In the Human-in-the-Loop approach ML solution to a task at hand is built in iterative process. Thus this process human can gain trust in the method.

Objective data vs metadata is a distinction which I proposed and it seems well suited for HITL. The pipeline which I presented is suited for ML engineers.

Intelligible eXplainable AI framework

InXAI: Stability

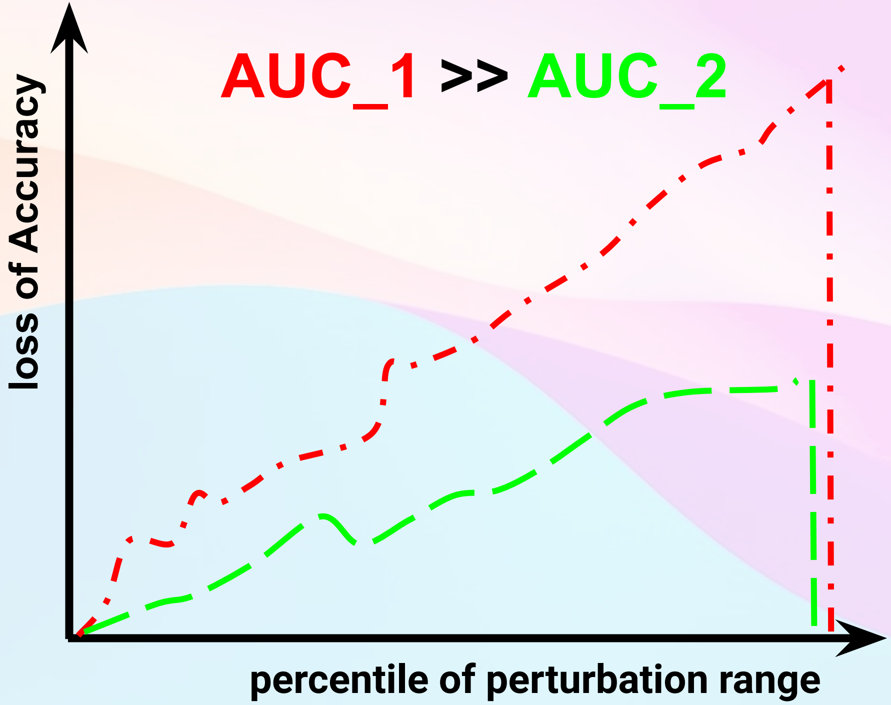


For **given explainer**, are explanations similar for similar input, measured with local Lipschitz continuity in the fixed neighborhood of any datapoint

Can one trust this explainer?

Bobek S., Mozolewski M., Nalepa G.J. (2021) Explanation-Driven Model Stacking. In: Paszynski M., Kranzlmüller D., Krzhizhanovskaya V.V., Dongarra J. J., Sloot P.M.A. (eds) Computational Science – ICCS 2021. ICCS 2021. Lecture Notes in Computer Science, vol 12747. Springer, Cham.

InXAI: Perturbational Accuracy Loss

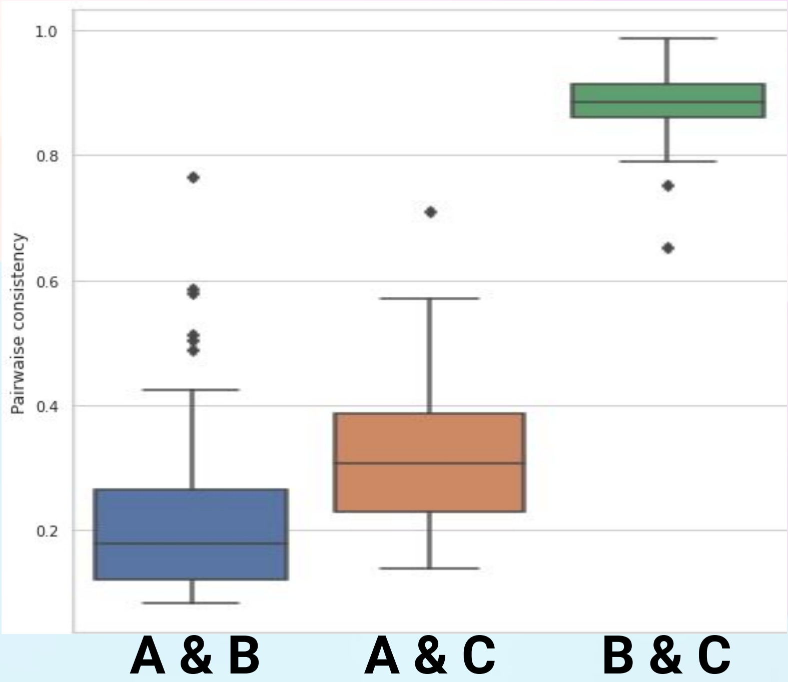


For **given explainer**, how accuracy deteriorates as the data get progressively perturbed, according to their inverse importance in explanation

Which explainer is the most accurate (in line with the ML model)?

Bobek S., Mozolewski M., Nalepa G.J. (2021) Explanation-Driven Model Stacking. In: Paszynski M., Kranzlmüller D., Krzhizhanovskaya V.V., Dongarra J. J., Sloot P.M.A. (eds) Computational Science – ICCS 2021. ICCS 2021. Lecture Notes in Computer Science, vol 12747. Springer, Cham.

InXAI: Consistency (pairwise)



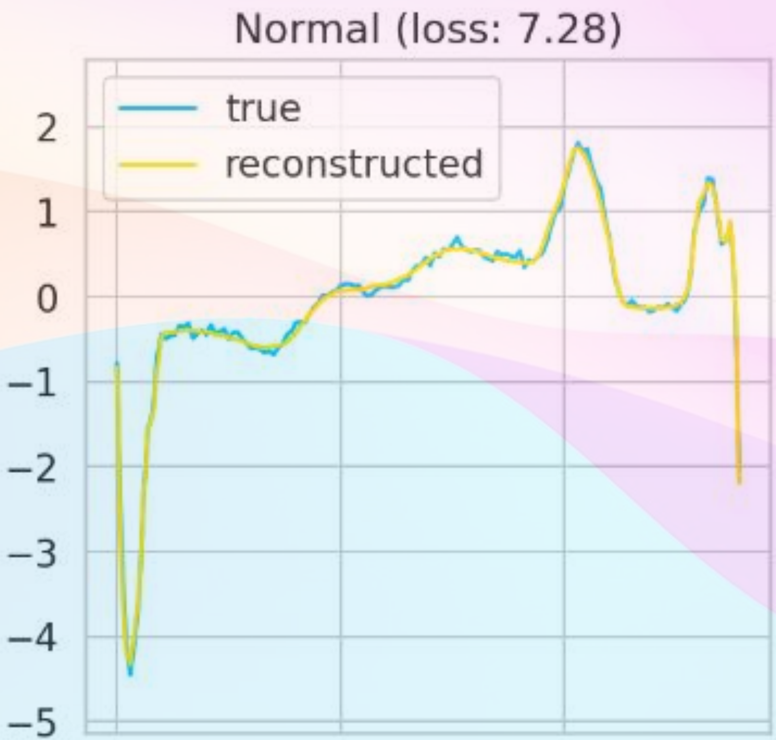
To what extent **different explainers** for predictions of ML model(s) are similar to each other (do agree)

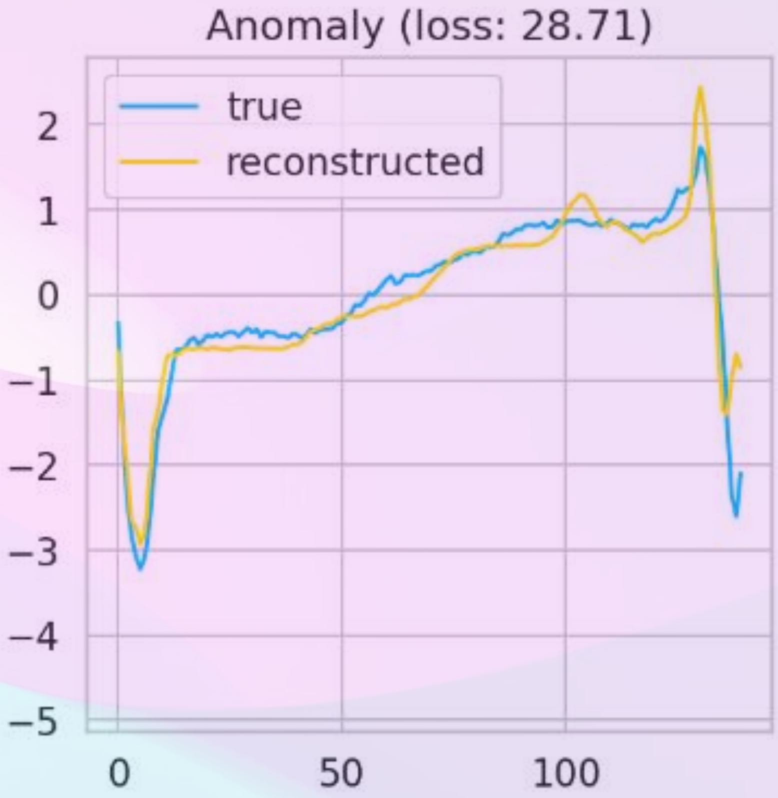
Can I exchange one explainer with another one?

Bobek S., Mozolewski M., Nalepa G.J. (2021) Explanation-Driven Model Stacking. In: Paszynski M., Kranzlmüller D., Krzhizhanovskaya V.V., Dongarra J. J., Sloot P.M.A. (eds) Computational Science – ICCS 2021. ICCS 2021. Lecture Notes in Computer Science, vol 12747. Springer, Cham.

WIP: Time Series

TS-data & anomaly detection with Autoencoders





InXAI Summary

Another way to build trust in XAI is to provide metrics that allow you to evaluate your explanations.

Metrics should answer questions such as:

1. Will I be able to trust the explanation in all circumstances? -> STABILITY
2. Which of the explanations is most consistent with the model being explained? -> AUC FOR PAC
3. To what extent do the different explanations agree with each other? Can I combine the explanations to make it even better? -> CONSISTENCY

Bibliography

1. Bobek S., Mozolewski M., Nalepa G.J. (2021) Explanation-Driven Model Stacking. In: Paszynski M., Kranzlmüller D., Krzhizhanovskaya V.V., Dongarra J. J., Sloot P.M.A. (eds) Computational Science – ICCS 2021. ICCS 2021. Lecture Notes in Computer Science, vol 12747. Springer, Cham. Download.

2. Mozolewski M., Jamshidi S., Bobek S., Nalepa G.J. (2022) Explain your clusters with words. The role of metadata in interactive clustering. CEUR Workshop Proceedings. Download.

Summary

XAI tailored for specific audience (agents/principals)

Trust in XAI:

- **via engagement of the user**
  - **HITL approach**
- **via metrics**