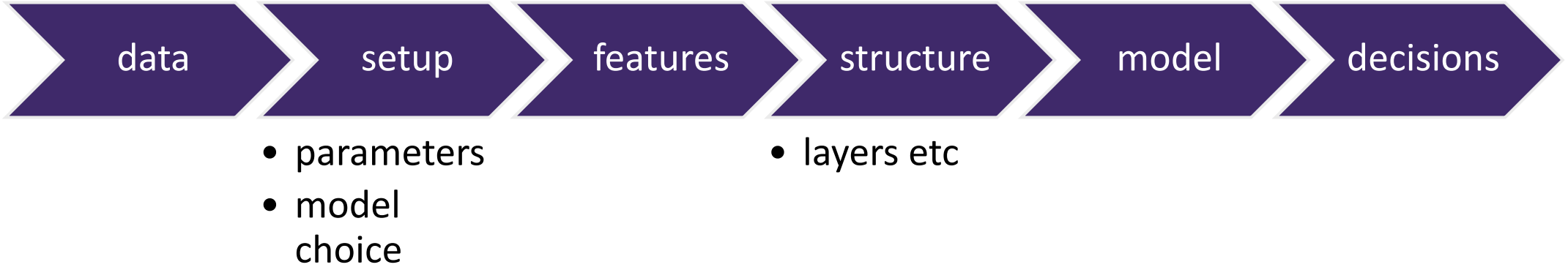
XAI MEETS NATURAL LANGUAGE PROCESSING

LARISSA HAAS

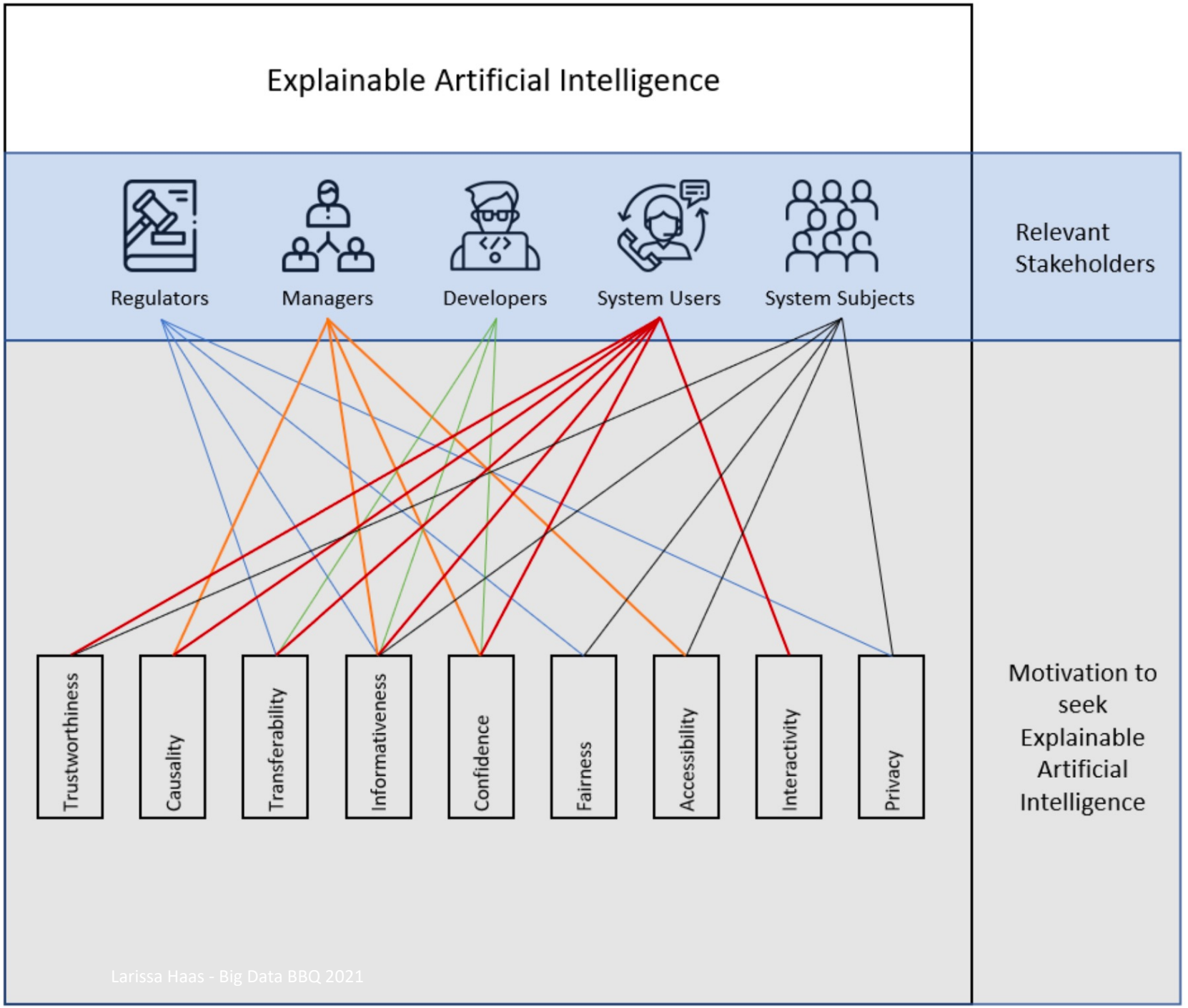
@L_R_SS_

Big Data BBQ 2021

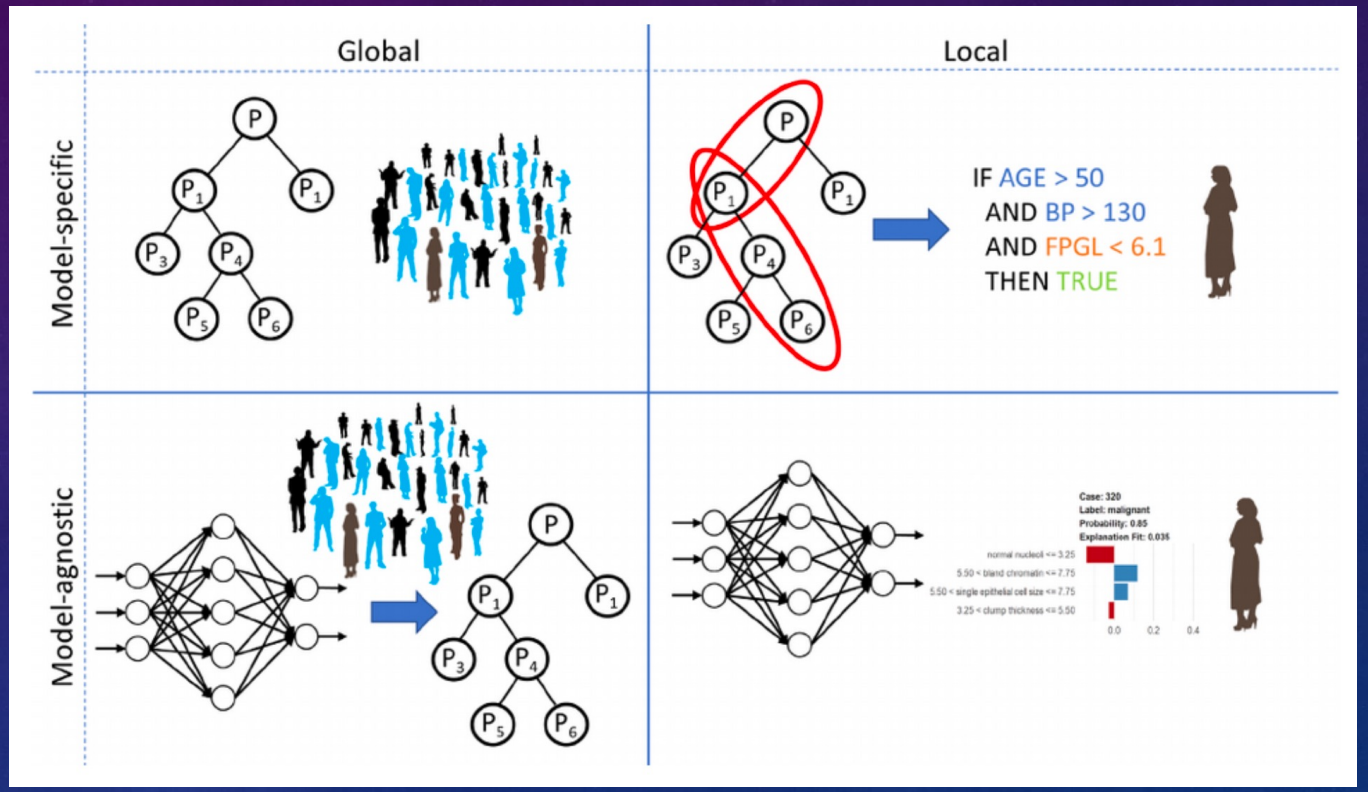
WHAT IS EXPLAINABLE AI?



WHY BOTHER?


Explainable Artificial Intelligence
Regulators
Managers
Developers
System Users
System Subjects
Relevant Stakeholders
Trustworthiness
Causality
Transferability
Informativeness
Confidence
Fairness
Accessibility
Interactivity
Privacy
Motivation to seek Explainable Artificial Intelligence

COMMON APPROACHES & EVALUATIONS

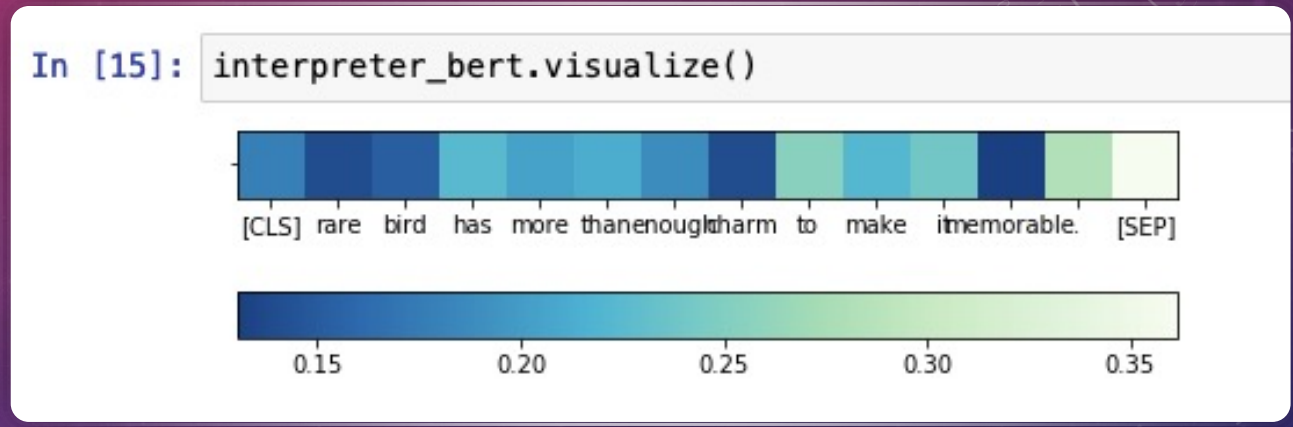


COMMON APPROACHES & EVALUATIONS

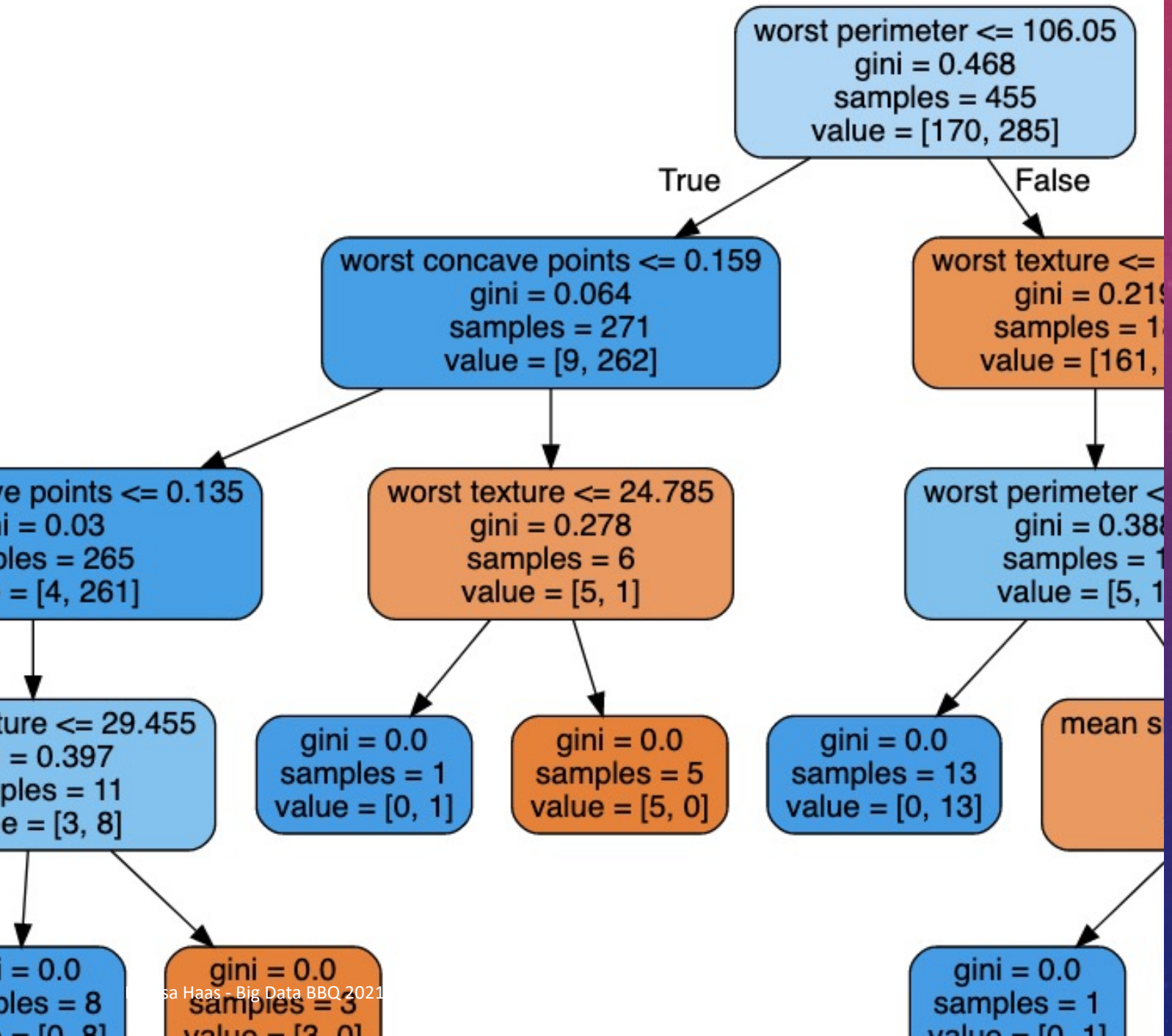
- human annotations
- feature importance and model structures
- understandable models
- correlations like LIME, SHAP, ELI5
- graphical representations

FEATURE IMPORTANCE & MODEL STRUCTURES

- use layers of BERT model to compute interpreter instances
- Sigma = the range that every word can change without changing **s** too much
- Sigma somewhat stands for the information loss of word $\mathbf{x}_i$ when it reaches **s**

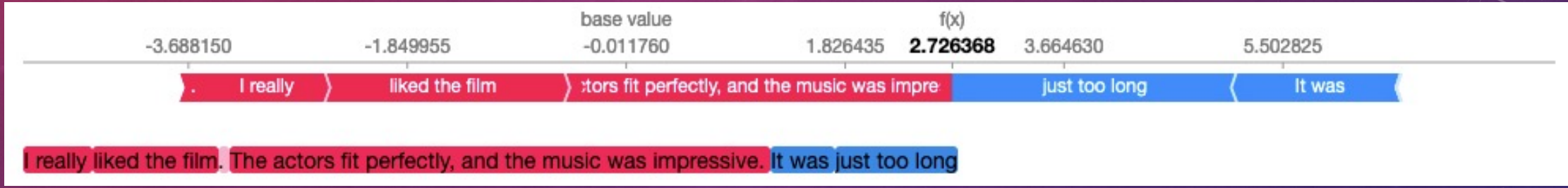


UNDERSTANDABLE MODELS



LIME, SHAP, ELI5

SHAP values



LIME: subset of SHAP, only locally
SHAP: locally and globally, but more costly

LIME, SHAP, ELI5

ELI5

```
In [31]: import eli5

         eli5.explain_prediction(gnb, datarows["text_new"][2741], vec=tfid, target_names=['fake', 'true'])
```

Out[31]: **y=fake** (score **-0.862**) top features

Contribution[?]	Feature
+0.632	Highlighted in text (sum)
+0.230	<BIAS>

reuters news agency declares war on trump in the most perfect way, trump humiliated donald trump s treatment of journalists is so bad that the reuters news agency is now advising reporters to treat trump like they would treat a foreign dictator.on tuesday, trump and his team announced that they will no longer send surrogates to appear on cnn, an escalation of the war trump has been waging against the news network, which he calls fake news because they won t promote his agenda.trump has even praised fox new s biased reporting as something cnn should copy.furthermore, trump and his team have repeatedly threatened and tried to intimidate journalists for doing their jobs just because they aren t writing puff pieces that kiss trump s ass.well, reuters is doing something about it.for 165 years, reuters has been bringing us news from around the world. they ve been in the most peaceful and democratic places, but they ve also been in war zones and reported on the most dangerous and tyrannical regimes in world history.in a message to staff on tuesday, reuters editor-in-chief steve adler advised his reporters to start dealing with trump the way they have dealt with brutal dictators in the past. it s not every day that a u.s. president calls journalists among the most dishonest human beings on earth or that his chief strategist dubs the media the opposition party, adler wrote. it s hardly surprising that the air is thick with questions and theories about how to cover the new administration. adler then revealed his solution for how reporters should handle trump.so what is the reuters answer? to oppose the administration? to appease it? to boycott its briefings? to use our platform to rally support for the media? all these ideas are out there, and they may be right for some news operations, but they don t make sense for

LIME, SHAP, ELI5

ELI5

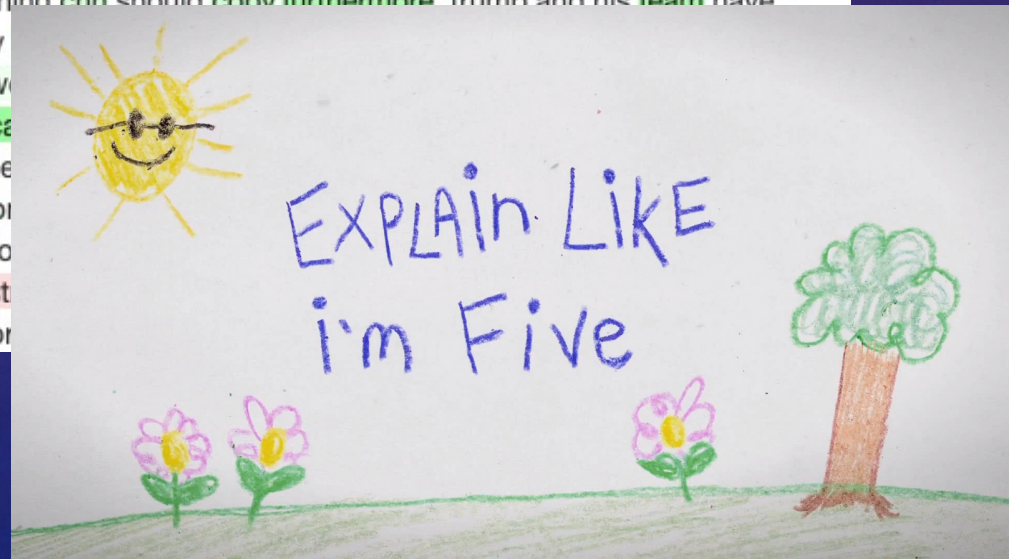
```
In [31]: import eli5

         eli5.explain_prediction(gnb, datarows["text_new"][2741], vec=tfid, target_names=['fake', 'true'])
```

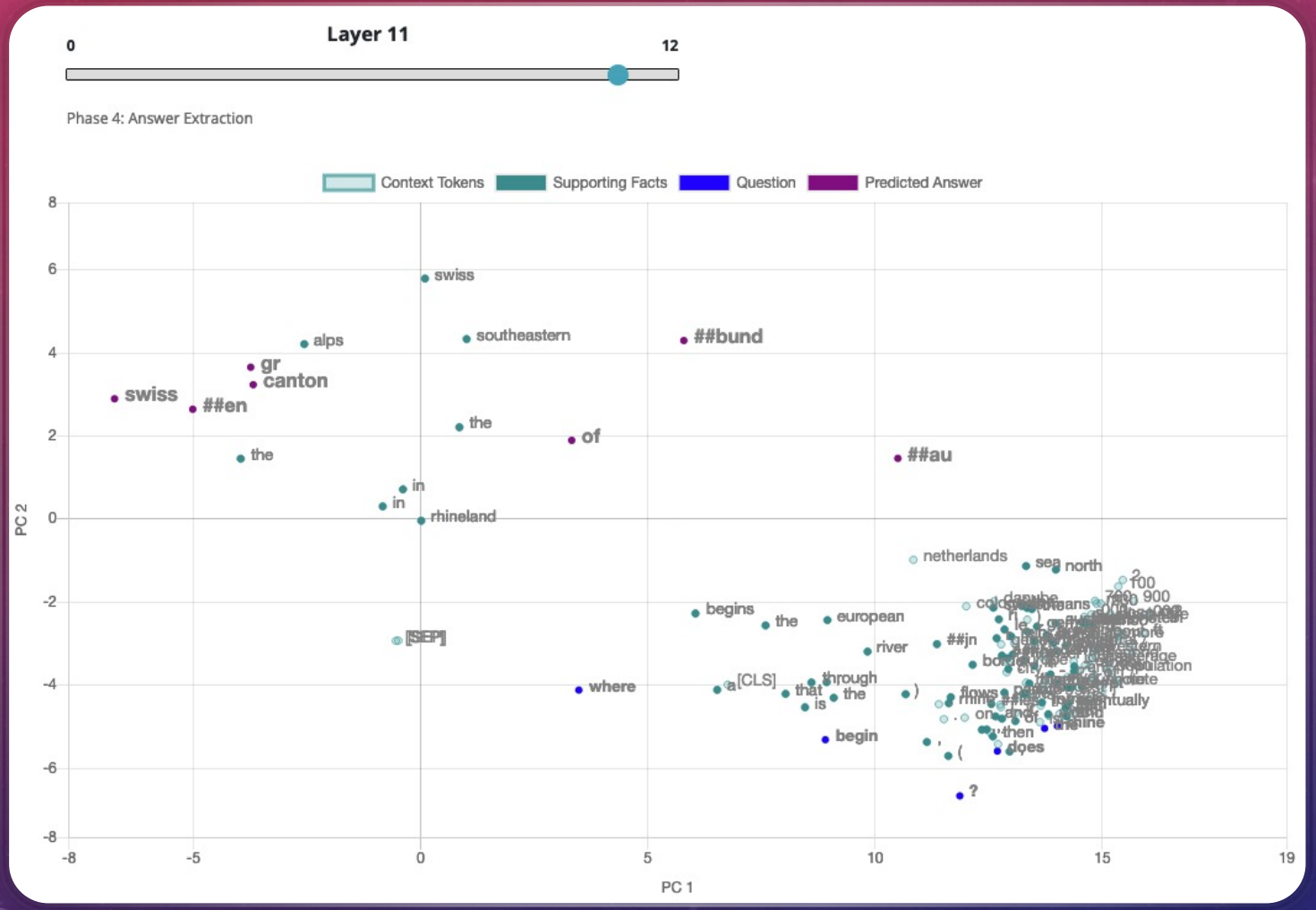
Out[31]: **y=fake** (score **-0.862**) top features

Contribution[?]	Feature
+0.632	Highlighted in text (sum)
+0.230	<BIAS>

reuters news agency declares war on trump in the most perfect way, trump humiliated donald trump s treatment of journalists is so bad that the reuters news agency is now advising reporters to treat trump like they would treat a foreign dictator.on tuesday, trump and his team announced that they will no longer send surrogates to appear on cnn, an escalation of the war trump has been waging against the news network, which he calls fake news because they won t promote his agenda.trump has even praised fox new s biased reporting as something cnn should copy furthermore, trump and his team have repeatedly threatened and tried to intimidate journalists for doing their jobs just because they doing something about it.for 165 years, reuters has been bringing us news from around the w places, but they ve also been in war zones and reported on the most dangerous and tyrannica reuters editor-in-chief steve adler advised his reporters to start dealing with trump the way the day that a u.s. president calls journalists among the most dishonest human beings on earth o adler wrote. it s hardly surprising that the air is thick with questions and theories about how to for how reporters should handle trump.so what is the reuters answer? to oppose the administ platform to rally support for the media? all these ideas are out there, and they may be right fo



GRAPHICAL REPRESENTATIONS



- good for word embeddings
- but only with dimensionality reduction

COMMON APPROACHES & EVALUATIONS

- human annotations
- feature importance and model structures
- understandable models
- correlations like LIME, SHAP, ELI5
- graphical representations

COMMON APPROACHES & EVALUATIONS

approach	evaluation for NLP
human annotations	for all kinds of applications difficult
Feature importances & model structures	depends on model
understandable models & feature importances	depending on preprocessing: but mostly not applicable, except as shadow model
correlations (LIME, SHAP, ELI5)	depending on preprocessing & model: can work quite well downside: slow and without feature dependence
graphical representations	good e.g. for evaluating embeddings / word vectors, but also only via SVD / PCA / t-SNE

See also https://www.arxiv-vanity.com/papers/2106.07410/

DIFFICULTIES FOR NLP MODELS

- high number of dimensions
- high variability in models (BoW / embeddings / …)
- set of tokens (aka text) != variables
- transition from local to global explainability
- difficult to show graphically (bec. of dimensionality and complexity)
- joint effects and effects of absent words

LESSONS LEARNED & OPEN TO DOS

EXPLAIN THE EXPLAINABILITY

ABSENSE OF WORDS

LOCAL IMPACT VS GLOBAL IMPACT

FIND A FITTING GRAPHICAL REPRESENTATION

REDUCE COMPLEXITY

INCLUDE HELP TEXTS

LINKS & FURTHER READING

- https://volpato.io/articles/1907-nlp-xai.html
- https://www.arxiv-vanity.com/papers/2106.07410/
- https://analyticsindiamag.com/hands-on-guide-to-interpret-machine-learning-with-shap/
- https://github.com/RajkumarGalaxy/StructuredData/blob/master/Interpret_ML_with_SHAP.ipynb
- https://christophm.github.io/interpretable-ml-book/shap.html
- https://shap.readthedocs.io/en/latest/index.html
- https://eli5.readthedocs.io/en/latest/overview.html
- https://mapmeld.medium.com/deciphering-explainable-ai-with-eli5-22c90a06a32a
- https://towardsdatascience.com/visualizing-word-embedding-with-pca-and-t-sne-961a692509f5
- https://microsoft.github.io/nlp-recipes/examples/model_explainability/
- https://visbert.demo.datexis.com/
- Alejandro Barredo Arrieta, Natalia D´ıaz-Rodr´ıguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador Garc´ıa, Sergio Gil- Lo´pez, Daniel Molina, Richard Benjamins, et al. Explainable artificial in- telligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information Fusion*, 58:82–115, 2020.
- Gregor Stiglic, Primoz Kocbek, Nino Fijacko, Marinka Zitnik, Katrien Ver- bert, and Leona Cilar. Interpretability of machine learning-based prediction models in healthcare. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 10(5):e1379, 2020.

THANKS!

- GitHub https://github.com/LarissaHa
- Twitter @l_r_ss_
- LinkedIn