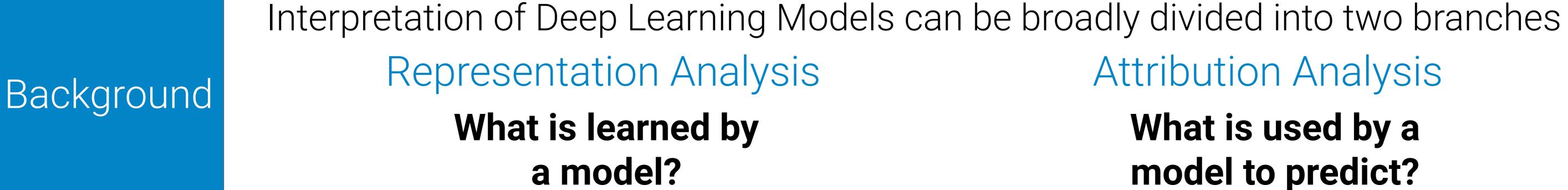
## NxPlain

## A Web-based Tool for Discovery of Latent Concepts Fahim Dalvi Nadir Durrani Hassan Sajjad Mus'ab Husaini Ummar Abbas Tamim Jabban {faimaduddin,ndurrani}@hbku.edu.qa



Goal

Overview

Provide a holistic view into the model by combining representation and attribution analysis

Leverage Latent Concept Analysis [1], Concept Alignment [2] and Attribution Analysis [3] to explain what knowledge a model has learned and how does it use the knowledge to make predictions

immigration helping takeover representation establishing globalists attacking genocide immigrant undocumented asylum criminals aliens coming resources advantage invasion hostile destroy taking advantage alien vicious invaders sell invaders invaders invaders invaders invaders invaders invaders invaders invaders seeking crossing terrorism enter illegals control nfiltrate steal entering control destroying entry successfully invaded

Terms used in hate-speech

Concept made up

of numbers

10,000

**93** 83

snail-like sleep-inducing squirm-inducing silly-looking stomach-turning phoney-feeling out-sized univac-like ugly-looking oking cheap-looking crummy-looking

Syntactic concept of



Concepts ^

against immigration policies

hyphenated words

Mannheim Baden-Baden Baden-W\u00fcrttemberg 105<sup>89</sup>94 Schleswig-Holstein Heidelberg Freiburg Trier Stuttgart **Prussian** Pomerania Bremen Armstadt Brandenburg Mainz Weimar Mecklenburg-Western Essen Hanove 2,500

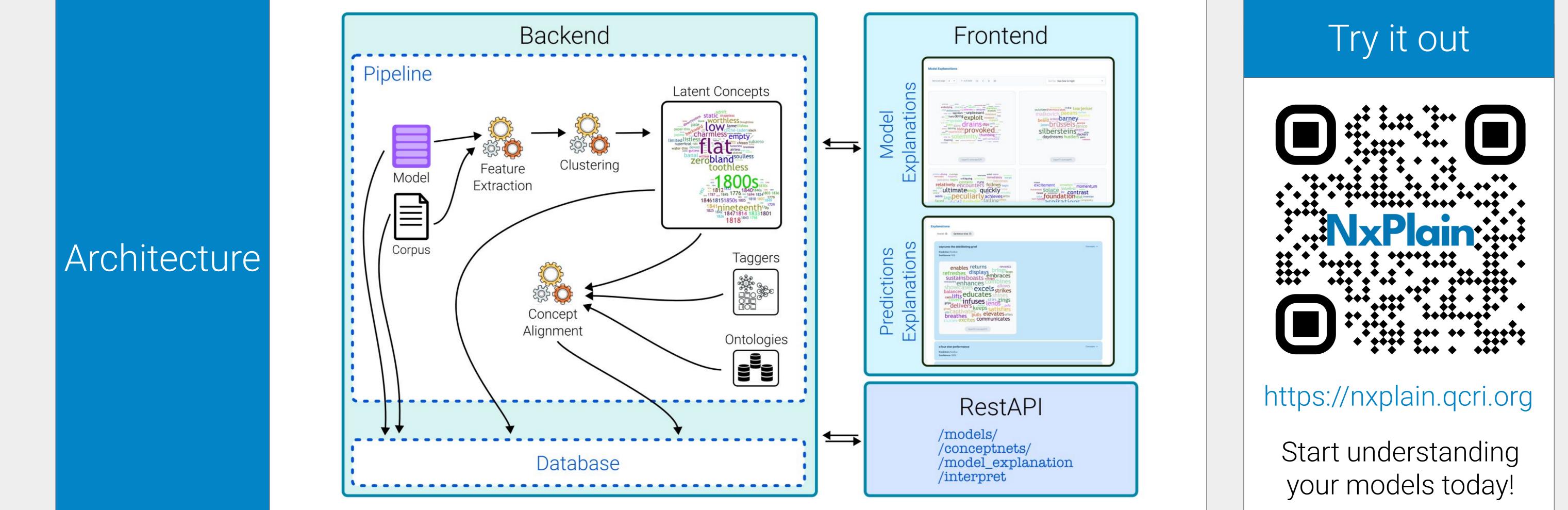
> Named entities in Germany

stirs delivers sits lends presents boasts casts excels trikes gives grips educates shines nines encourages ZINGS creates showcases combines leaps sustains<sup>puts</sup> captivates buoy elevates excites breathes

[1] Dalvi, Fahim, et al. "Discovering Latent Concepts Learned in BERT." International Conference on Learning Representations 2022

[2] Sajjad, Hassan, et al. "Analyzing Encoded Concepts in Transformer Language Models." North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2022

[3] Sundararajan, Mukund, Ankur Taly, and Qiqi Yan. "Axiomatic attribution for deep networks." International conference on machine learning. PMLR, 2017





https://neurox.qcri.org

