Monotonic Representation of Numeric Properties in Language Models

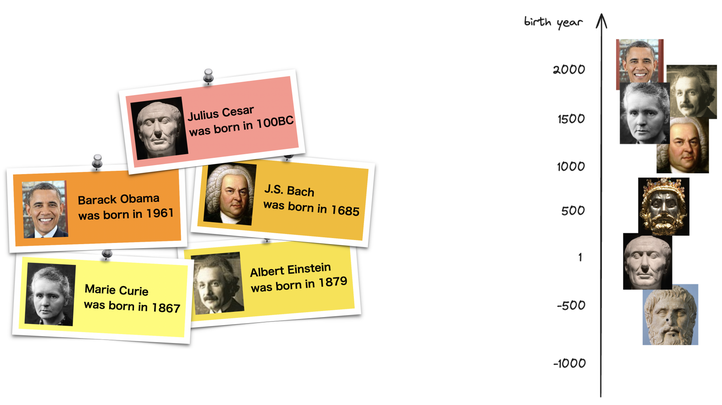
Benjamin Heinzerling[1, 2] Kentaro Inui[3, 2, 1]

[1]RIKEN AIP [2]Tohoku University [3]MBZUAI

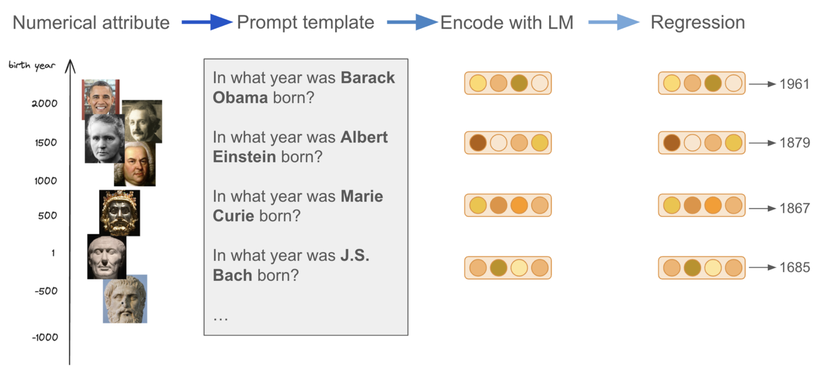
Overview

- LMs can answer questions involving numeric properties
- For example "When was Karl Popper born?"
- How is this knowledge represented inside the LM?
- We find *monotonic representations* of numeric properties
- There are directions in activation space that correlate with the value of numeric properties
- By activation patching along such directions, LM output changes accordingly

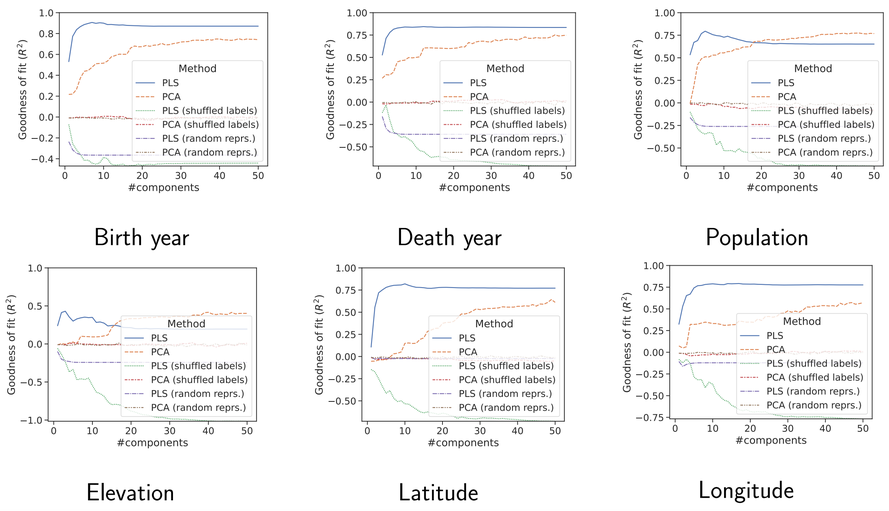
Are entity representations unordered or ordered?



Finding property-encoding directions via PLS regression

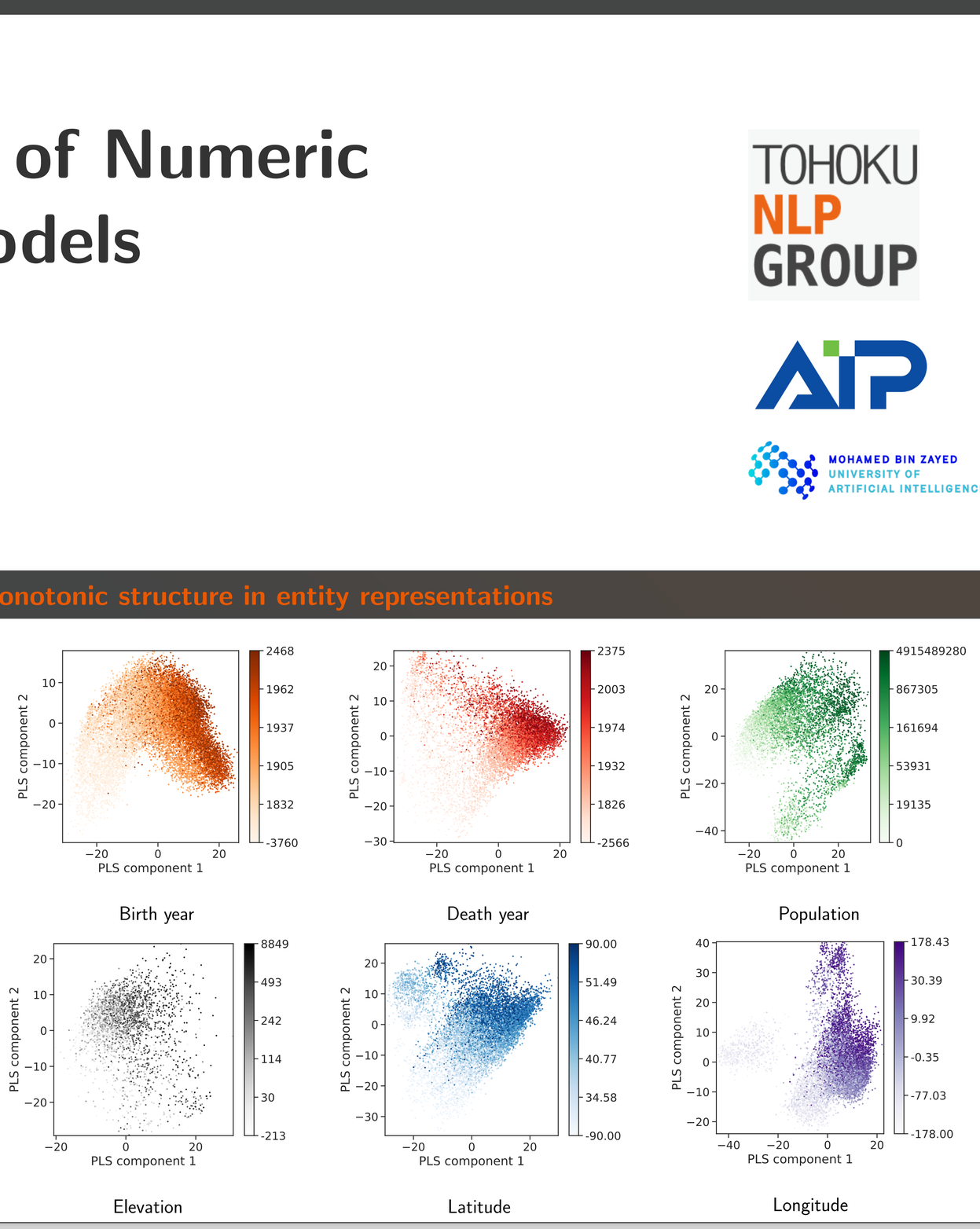


Low-dimensional subspaces predict numeric properties

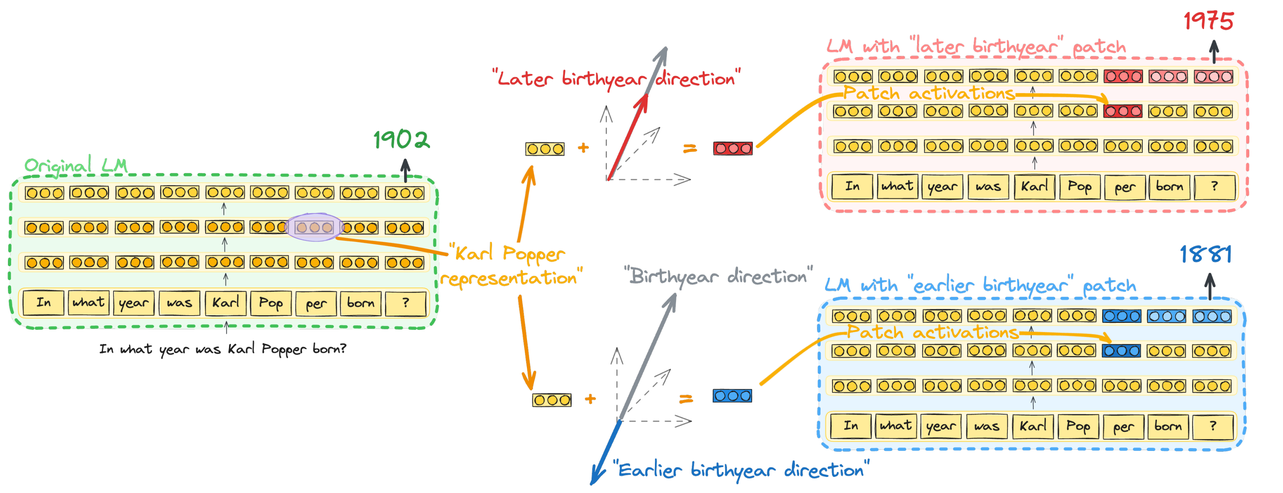


Side effects

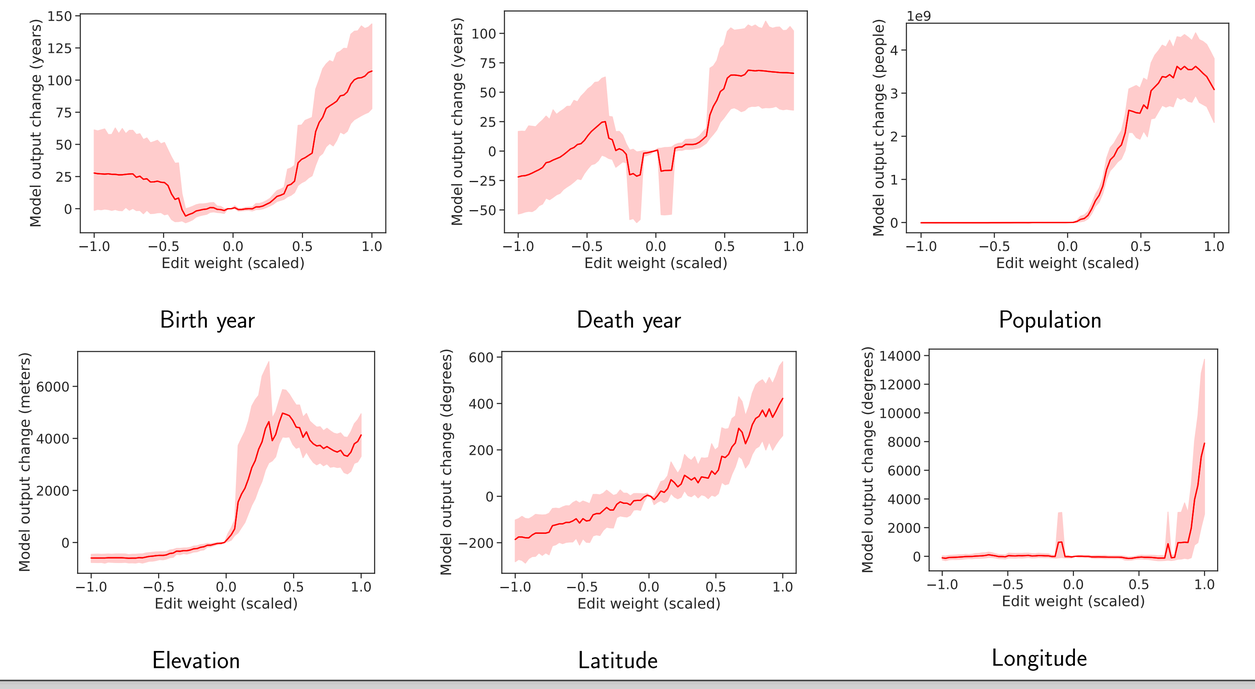
Monotonic structure in entity representations



Does representation affect computation?



Effect of editing representations along property-encoding directions



Data sample

Property	Prop. ID	Entity	Entity ID	Prompt	Value	Unit
birthyear	P569	Nina Foch	Q235632	In what year was Nina Foch born?	1924	annum
death year	P570	Johannes R. Becher	Q58057	In what year did Johannes R. Becher die?	1958	annum
population	P1082	Akhisar	Q209905	What is the population of Akhisar?	173026	1
evelation	P2044	Sondrio	Q6274	How high is Sondrio?	360	metre
longitude	P625.long	Korean Empire	Q28233	What is the longitude of Korean Empire?	126.98	degree
latitude	P625.lat	Küsnacht	Q69216	What is the latitude of Küsnacht?	47.32	degree