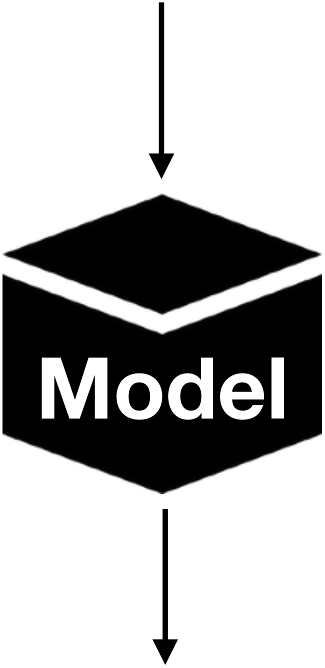
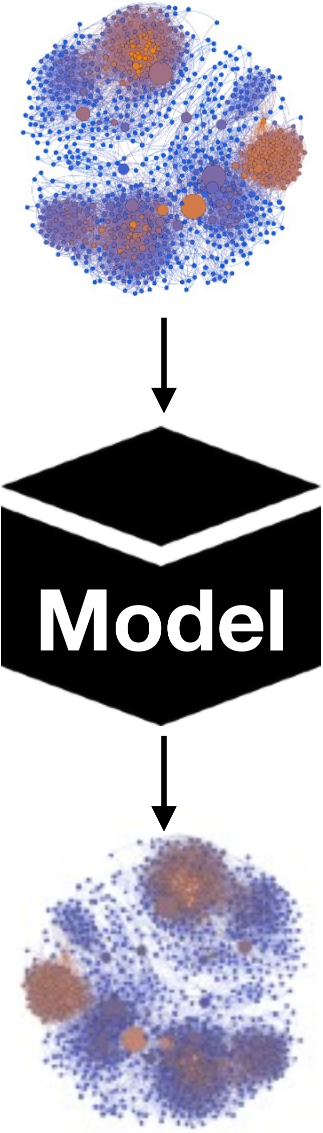
Explaining
Structured Models

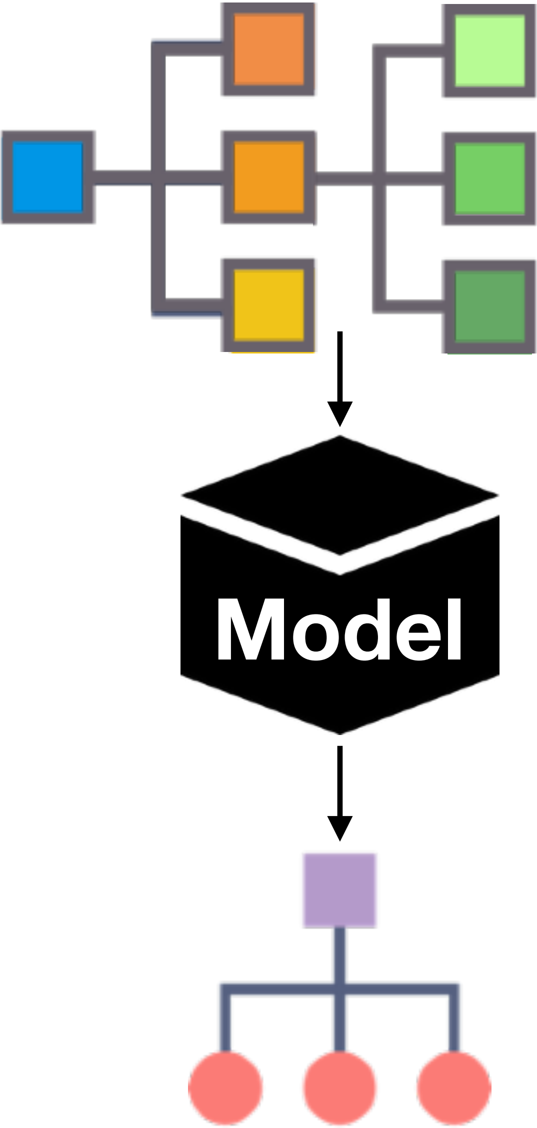
*"Mary did not slap
the green witch"*



Model

*"Mary hat die grüne Hexe
nicht geschlagen"*


Model

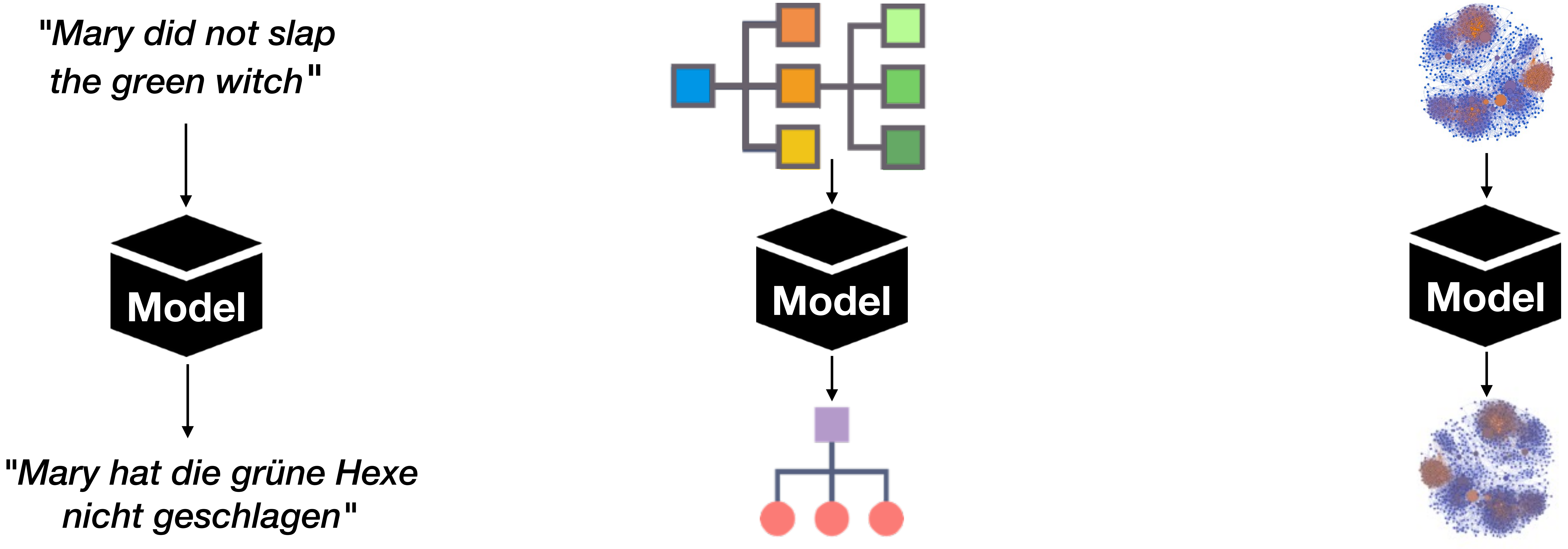

Model

- Structured inputs/outputs **vary in size and complexity**

- What parts do we explain?

- What does ***"local"*** mean for a structured input?

Explaining
Structured Models



"Mary did not slap the green witch"

"Mary hat die grüne Hexe nicht geschlagen"

- Structured inputs/outputs **vary in size and complexity**
- What parts do we explain?
- What does ***"local"*** mean for a structured input?

Interpretability
for Black-Box Seq2Seq Models

AM+ Jaakkola, *EMNLP'17*