

# ANITI

Communauté  
d'universités  
et établissements  
de Toulouse



ANITI Days 2026:  
Parole aux PhD et PostDocs!

# Language Models' Interpretability

## Antonin Poché

Supervisors: Pr. N. Asher, Pr. P. Muller, and Dr F. Jourdan



# ANITI Concept-based explanations: What?

- Examples with the BIOS dataset:
- assign occupation to biographies

Nurse

Surgeon

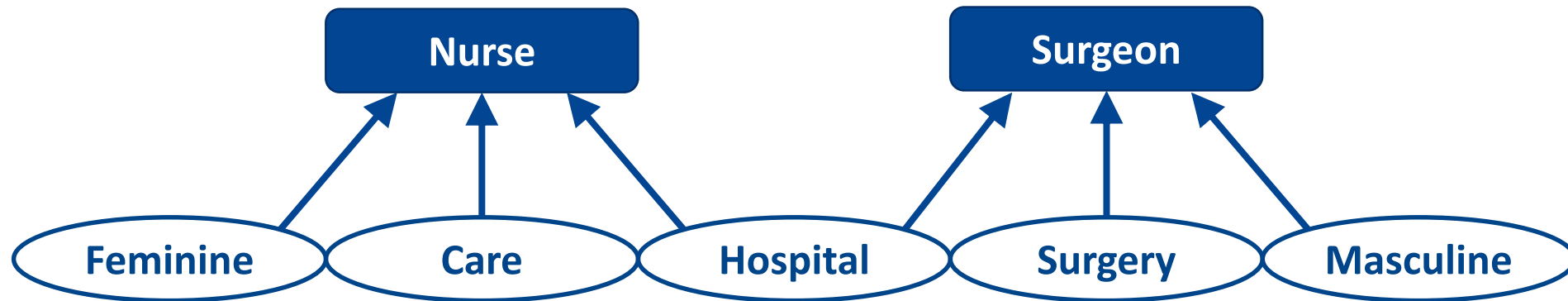
# ANITI Concept-based explanations: What?

- Examples with the BIOS dataset:
- assign occupation to biographies



# ANITI Concept-based explanations: What?

- Examples with the BIOS dataset:
- assign occupation to biographies



# ANITI Concept-based explanations: What?

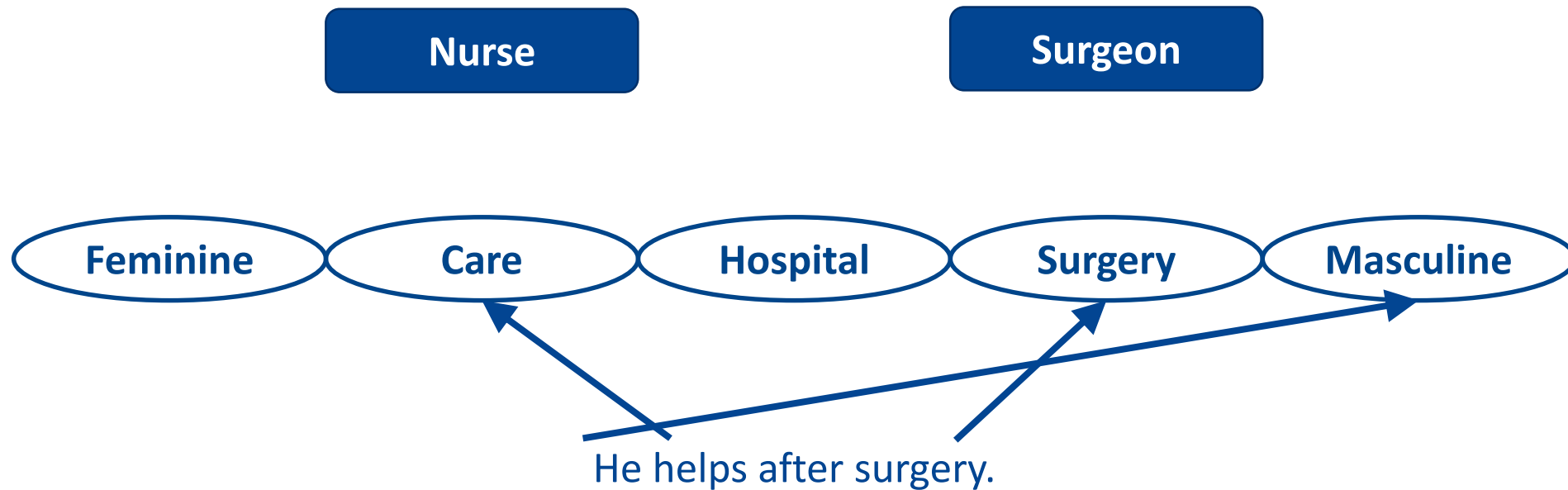
- Examples with the BIOS dataset:
- assign occupation to biographies



He helps after surgery.

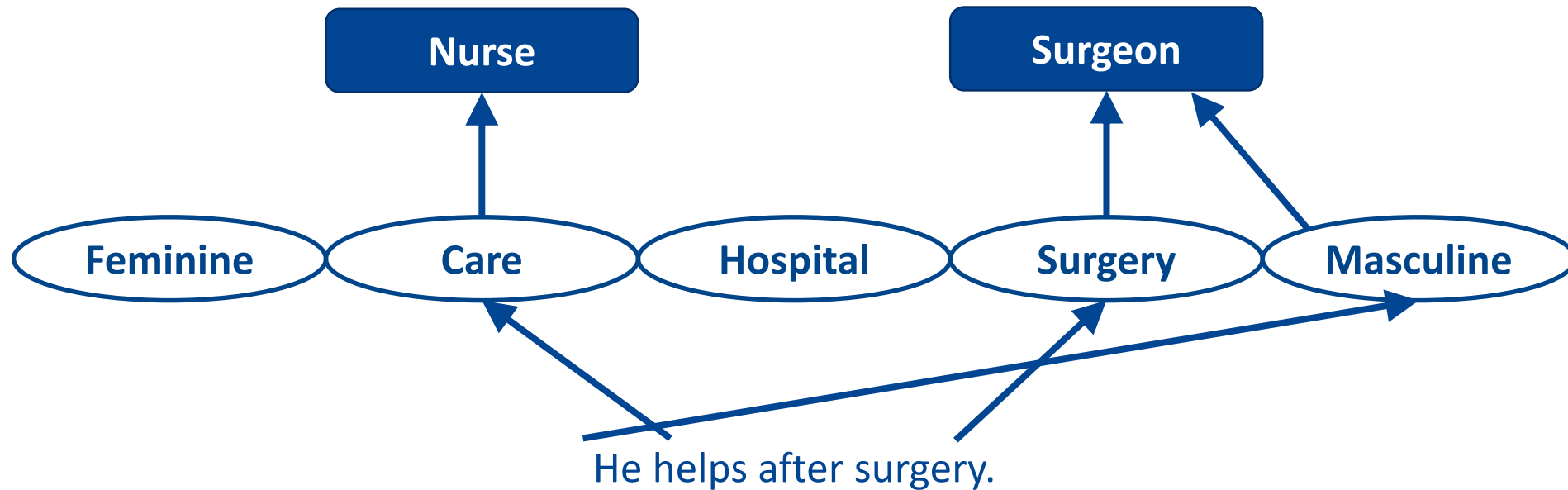
# ANITI Concept-based explanations: What?

- Examples with the BIOS dataset:
- assign occupation to biographies

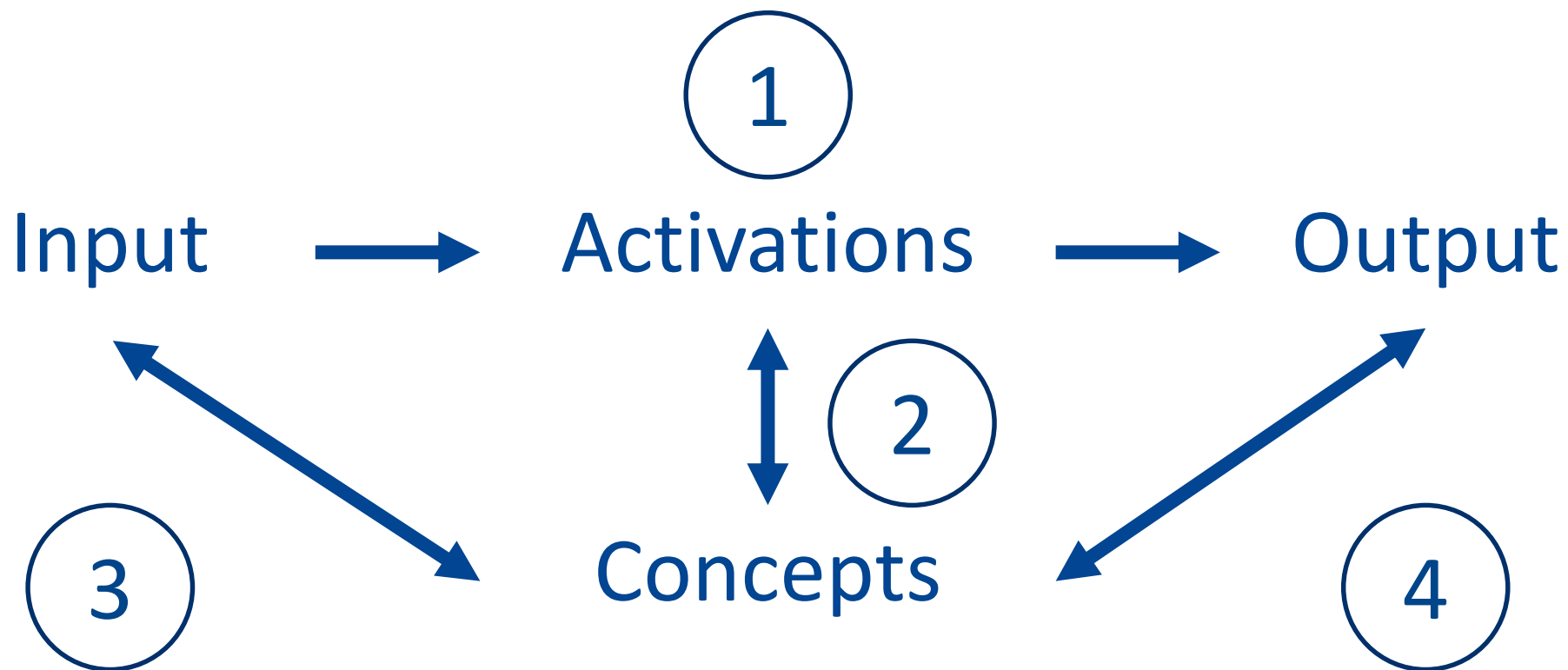


# ANITI Concept-based explanations: What?

- Examples with the BIOS dataset:
- assign occupation to biographies



# ANITI Concept-based explanations: How?





# Interpreto

