Interpretability for Bias Detection in Machine Translation

- NLP methods often incorporate **biases** present in training data
    - Gender ↔ occupation stereotypes [Caliskan et al. 2017]
    - Sexist adjective associations [Bolukbasi et al. 2016]

- Can we use our framework to **understand** these biases?

- **Black-box:** Azure's MT service, English ⟶ French

- **Inputs:** sentences containing bias-prone words

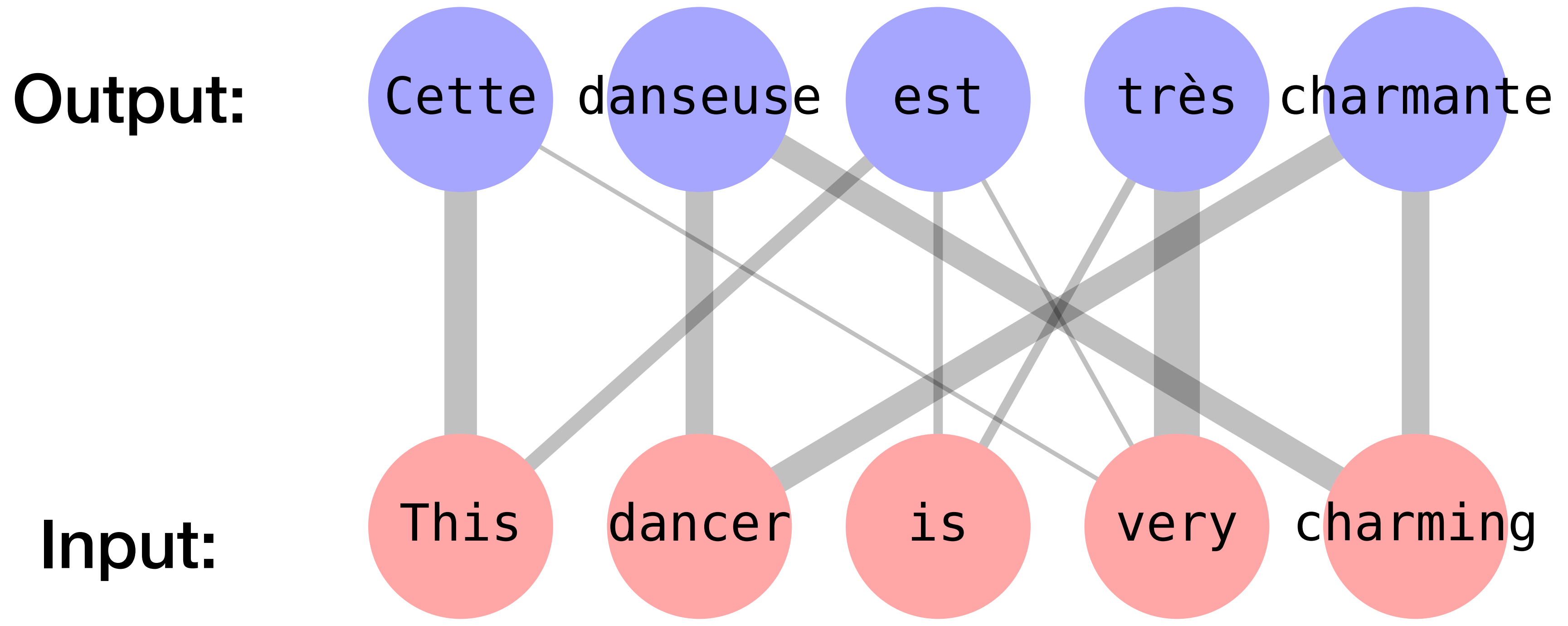
- **Results:** model exhibits strong grammatical gender preferences
  - *doctor, professor, smart, talented -> translates to* ***masculine***
  - *dancer, nurse, charming, compassionate -> translates to* ***feminine***

Interpretability for Bias Detection in Machine Translation

- NLP methods often incorporate **biases** present in training data
  - Gender ↔ occupation stereotypes [Caliskan et al. 2017]
  - Sexist adjective associations [Bolukbasi et al. 2016]

- Can we use our framework to **understand** these biases?
- **Black-box:** Azure's MT service, English ⟶ French
- **Inputs:** sentences containing bias-prone words
- **Results:** model exhibits strong grammatical gender preferences
  - *doctor, professor, smart, talented -> translates to* ***masculine***
  - *dancer, nurse, charming, compassionate -> translates to* ***feminine***

Interpretability for Bias Detection in Machine Translation


Output:
Cette
danseuse
est
très
charmante
Input:
This
dancer
is
very
charming