Using NLP to Improve Quality of Occupational Injury and Illness Data

Nathan Bartley
Alex Measure
August 4th, 2021

Death Certificates

- Arguably most important source document
  - Frequently prepared by trained individual
  - Comprehensive
- Challenges
  - No access in national office
  - Every state does it differently
  - Frequently done on paper

What We've Done

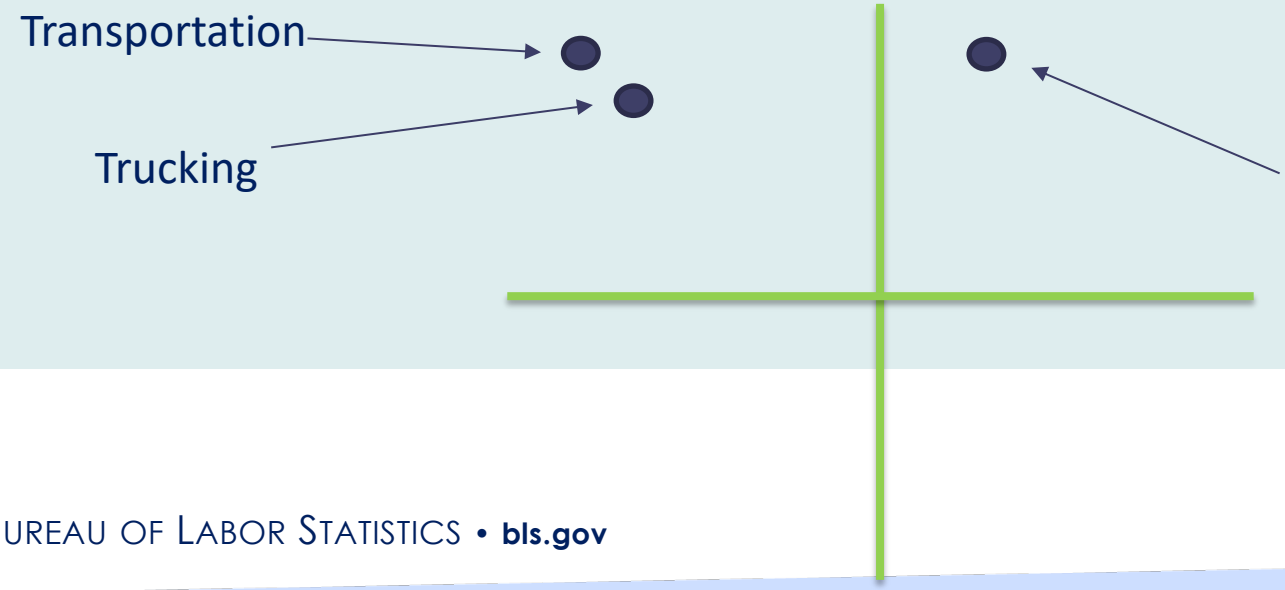
- Access for 3 states
- Built programs to normalize fields
  - Convert columns with similar data to same name
  - Convert values to comparable formats
    - (i.e. date might be 12-25-2021 or Dec 25 2021 or 2021/12/25)
- Compare to equivalent CFOI data
- Used in production already!

Results

- Quite a few typos
  - Dates off by a digit
  - Even some genders wrong
- What about trickier comparisons
  - Is occupation “Trucker” consistent with “Transportation”

Comparing language “meaning”

- Neural network trained on huge collections of text
- Learns vector representations of text
- Distance between vectors reflects difference in meaning



Long Term Plan