OSS in the Federal Statistical System
An Example from BLS

David H. Oh

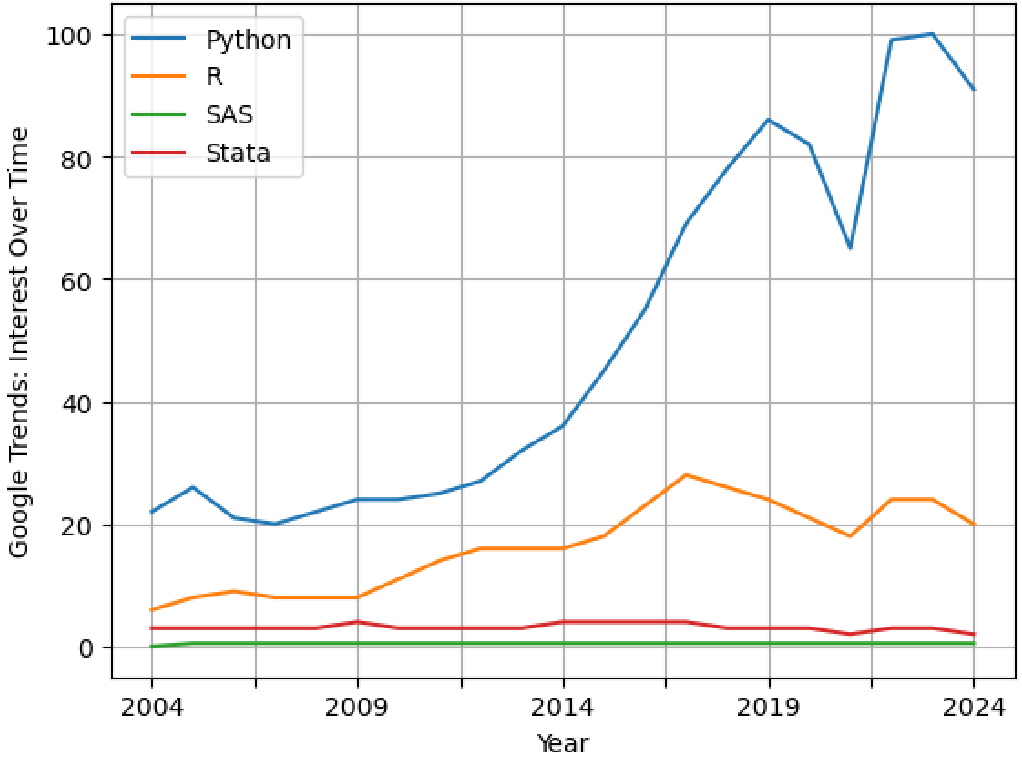
Supervisory Data Scientist

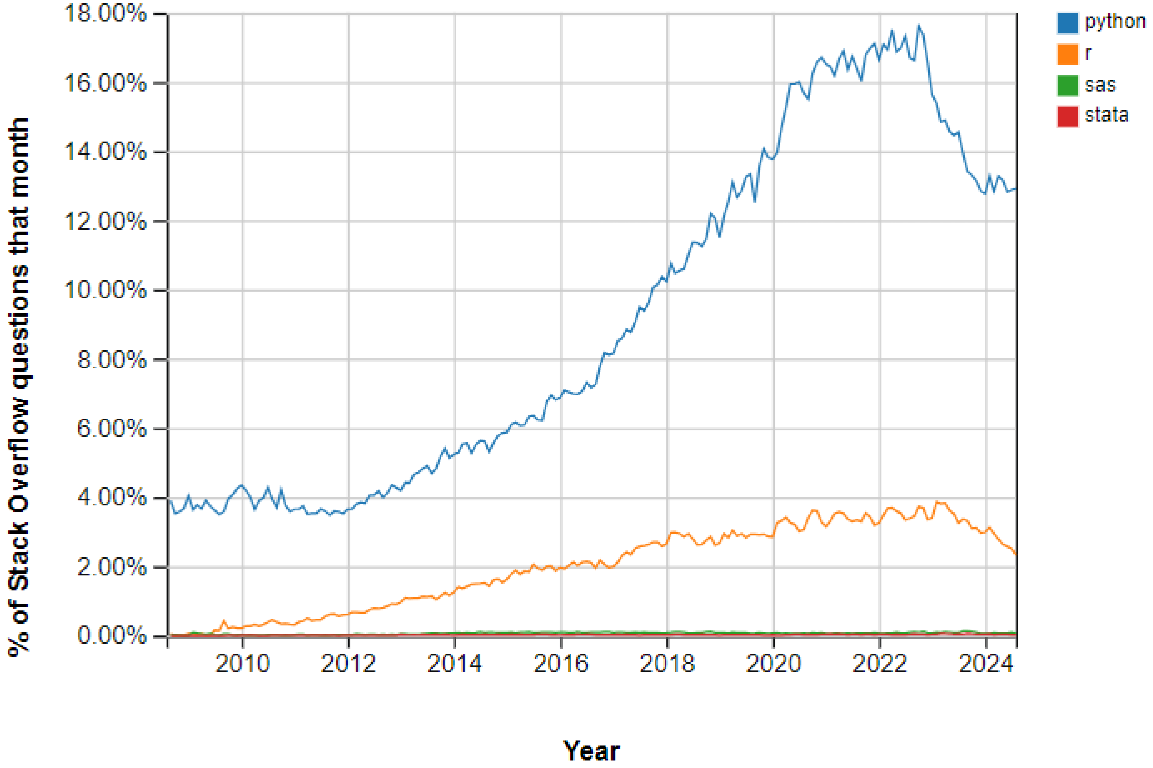
2024 FCSM Research & Policy Conference: Session B-3

October 22, 2024

Disclaimer: The views expressed herein are those of the author(s) and do not necessarily reflect those of the Federal Government, Department of Labor, or the Bureau of Labor Statistics. All results have been reviewed to ensure that no confidential information is disclosed.

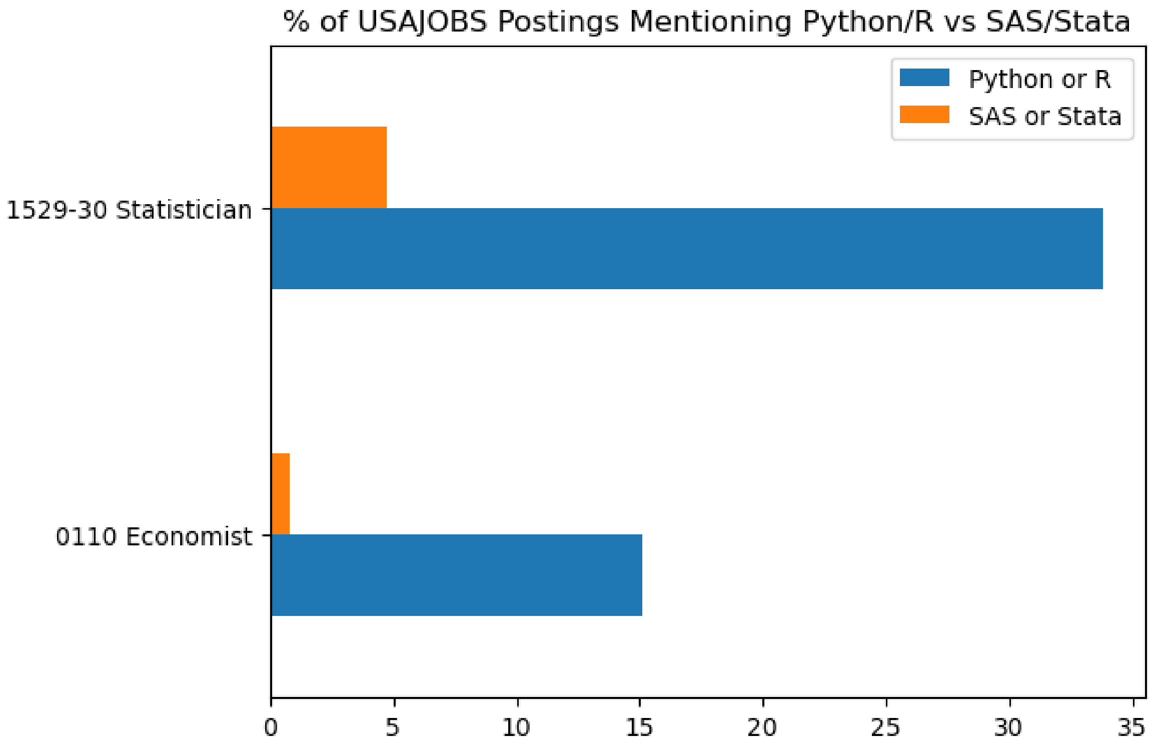
Growing Interest in Open-Source Software





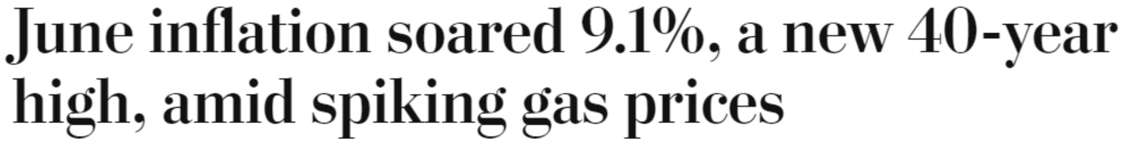
Growing Interest in Open-Source Software

- The growing interest and adoption of OSS is also evident within the federal government
- The recent job postings on USAJOBS for Statisticians (1530), Mathematical Statisticians (1529), and Economists (0110) are more likely to mention skills in OSS like Python or R over proprietary software like SAS or Stata

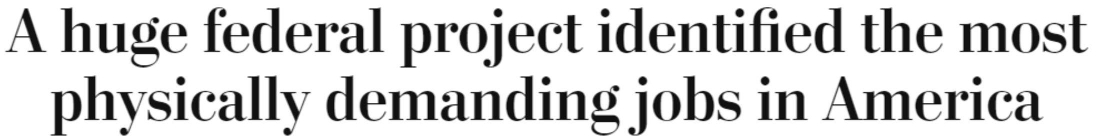


Bureau of Labor Statistics

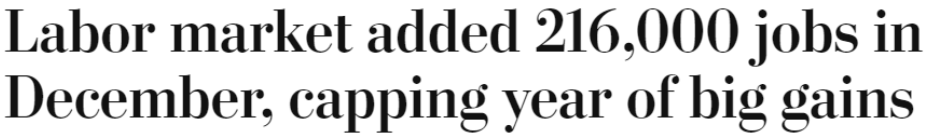
- The **Bureau of Labor Statistics (BLS)** is the principal fact-finding agency for the federal government in the broad area of labor economics and statistics.



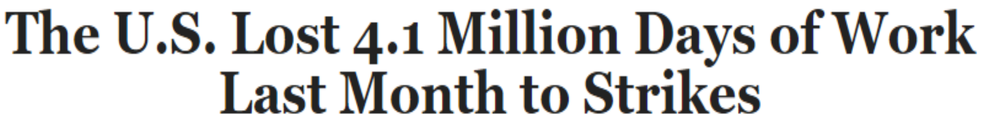
June inflation soared 9.1%, a new 40-year high, amid spiking gas prices



A huge federal project identified the most physically demanding jobs in America



Labor market added 216,000 jobs in December, capping year of big gains

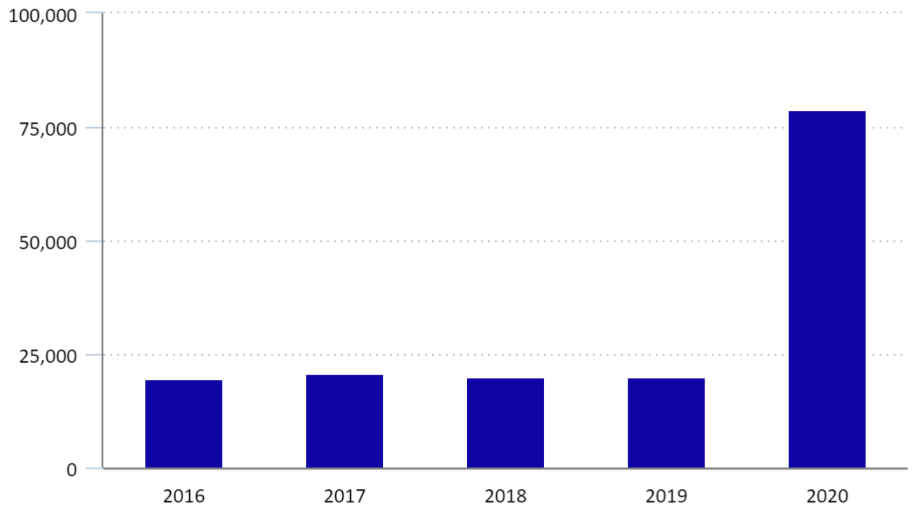


The U.S. Lost 4.1 Million Days of Work Last Month to Strikes

Survey of Occupational Injuries and Illnesses

- Annual establishment survey collecting injury and illness information since 1972
- Information collected:
  - Total number of cases resulting in days away from work or days of job transfer and work restrictions
  - Detailed case and demographic information about some injury or illness cases
- 200,000+ descriptions of work-related injuries and illnesses each year

Chart 1. Number of nonfatal occupational injury and illness cases involving days away from work, registered nurses, private industry, 2016–2020



Survey of Occupational Injuries and Illnesses

Example Narrative

Job title: sanitation worker

What was the employee doing just before the incident?
mopping floor in gym

What happened?
slipped on wet floor and fell

What part of the body was affected?
fractured right arm

What object directly harmed the employee?
wet floor

Codes Assigned

SOC: 37-2011 (Janitor)

OIICS-Nature: 124 (Fractures)

OIICS-Part: 420 (Arm)

OIICS-Event: 4312 (Fall, slipping)

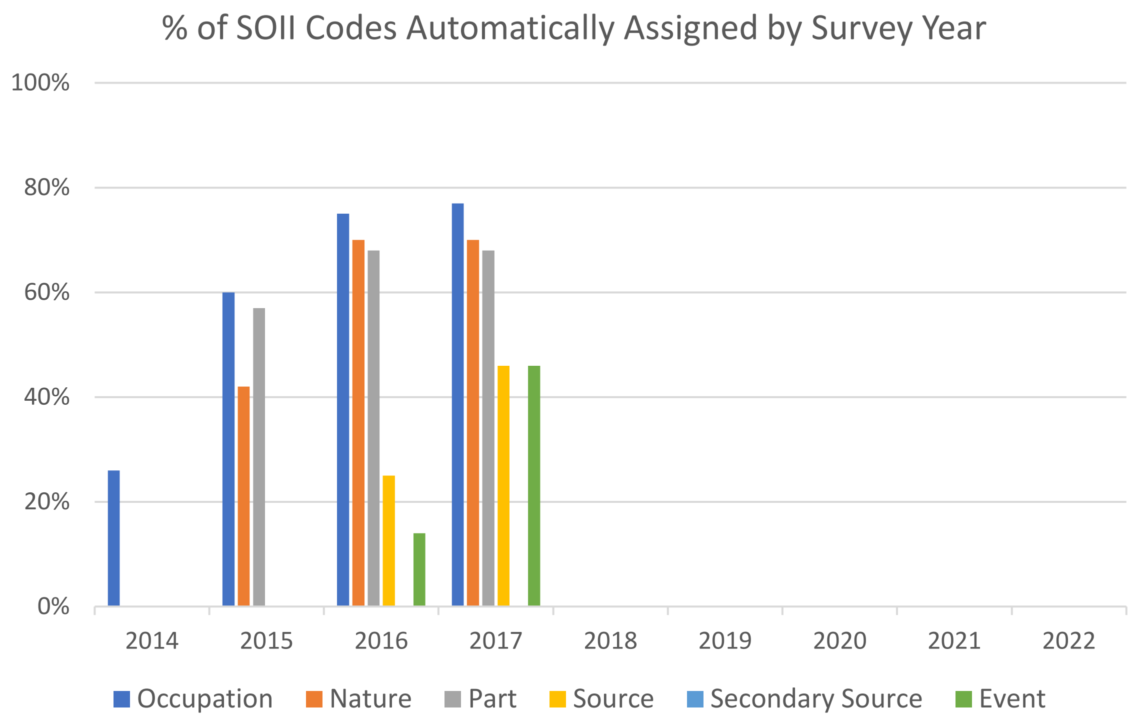
OIICS-Source: 6624 (Surface)

Computer-Assisted Coding (CAC)

- Began more than 10 years ago
- Manual coding was time and resource intensive
- People weren't coding consistently across regions
  - Two experts coding exact same narratives: ~70% agreement
- Can computers help?

SOII CAC Timeline

- The Past
  - Logistic Regression
  - Use of Python and its packages allowed a convenient way to process natural language
  - Initially used for review only
  - Usage expanded gradually over time

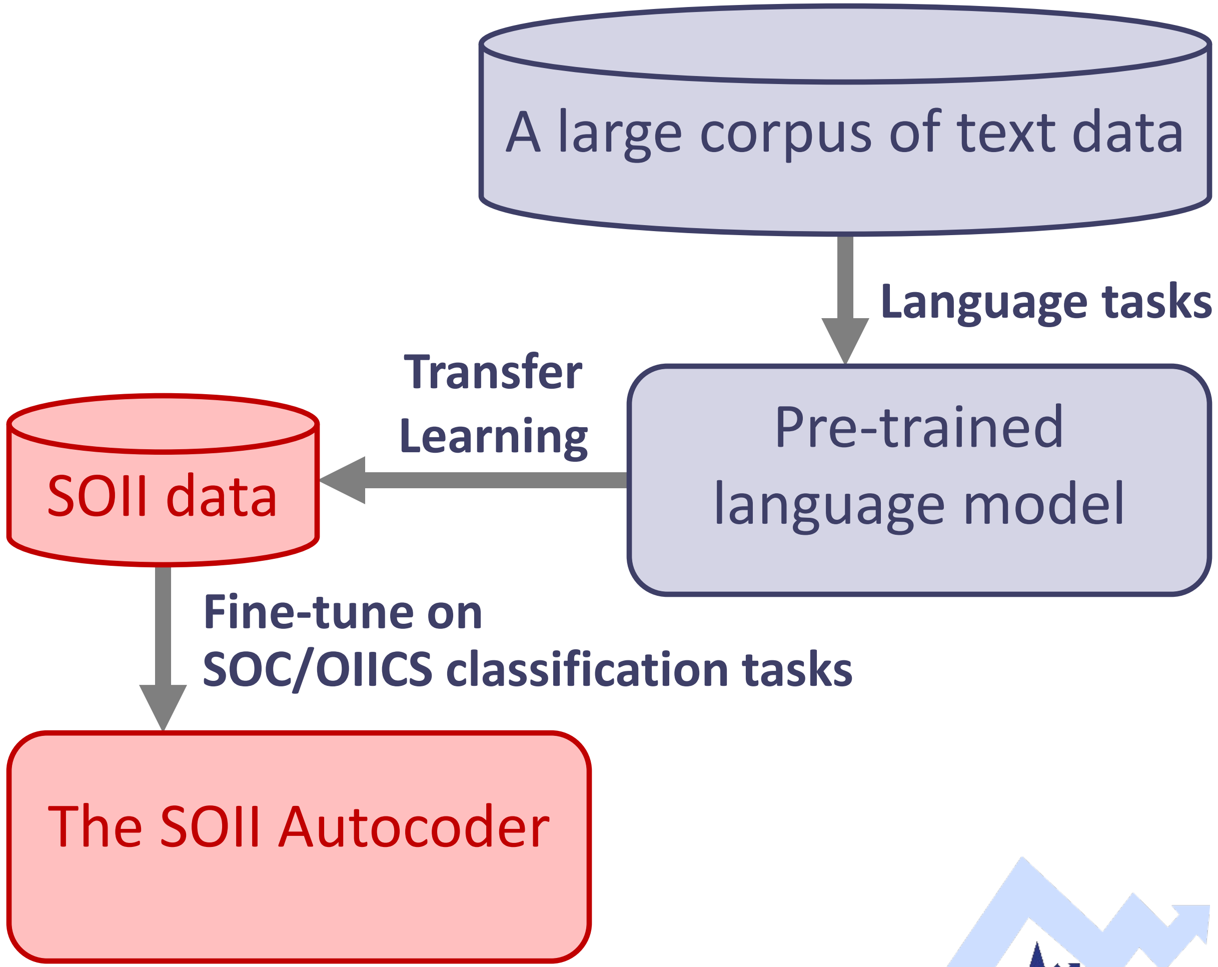


Past Challenges

- Context matters
  - “A man fell on a car” ≠ “A car fell on a man”
- Meaning of words
  - “lansaper” = “landscaper”
  - “eye” = “eyeball”
- Coding tasks are related

Solution: Neural Network

- Timeline:
  - 2018-2020: LSTM model
  - 2021: Transfer learning using transformer-based model
- Transfer learning leveraged an existing open-source language model



A large corpus of text data

Language tasks

Pre-trained language model

Transfer Learning

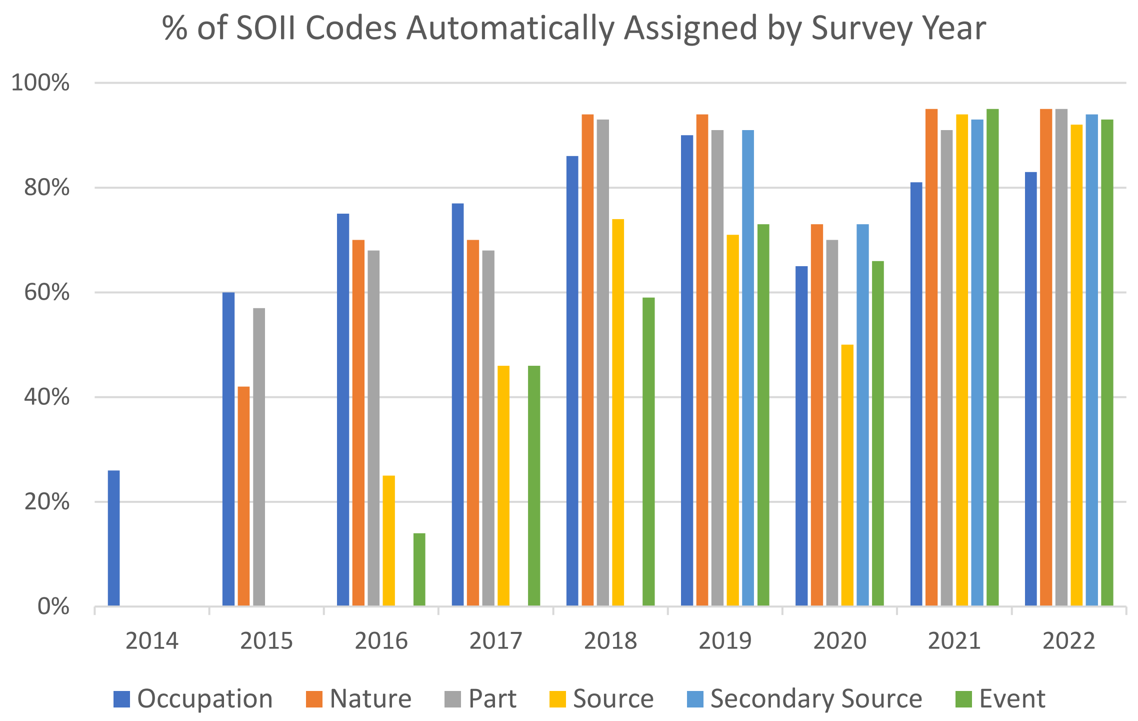
SOII data

Fine-tune on SOC/OIICS classification tasks

The SOII Autocoder

SOII CAC Timeline

- The Present
  - Neural Network architecture
    - LSTM
    - Transformer
  - Usage expanded to secondary source

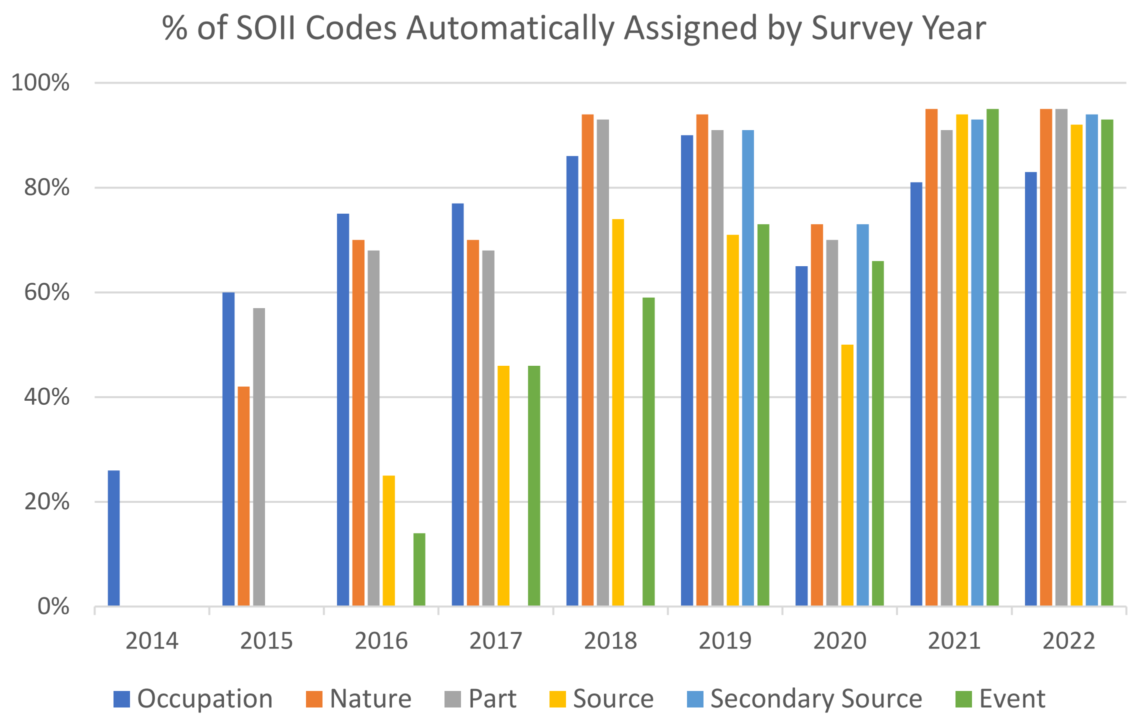


Present Challenges

- Accounting for potential biases inherited from the pre-trained models
  - Data used to train pre-trained models contain unfiltered content from the internet, which can be biased
- Fine-tuned model not shareable with external parties (yet)
  - Developed for internal use

SOII CAC Timeline

- The Future
  - Measure and account for inherited biases if it exist
  - Apply privacy-preserving mechanism
  - Practice responsible AI development and usage



Contact Information

David H. Oh
Supervisory Data Scientist

Office of Compensation and Working Conditions
202-691-5985
Oh.David@bls.gov