NLP in Practice:
Applying Natural Language Processing to Survey Text Data

Roundtable Discussion

FedCASIC
April 16, 2024

Natural Language Processing

Models that make use of human language as data

- Rule-based or machine learning approaches
- Tasks such as
  - Speech recognition and text to speech
  - **Document classification**
  - Summarization of documents
  - Information retrieval
  - Topic modeling
  - Named entity recognition
  - **Text matching**
  - Machine translation
  - And *many* more

Agenda

- Introductions
- Q&A with the Panel Moderator
- Q&A with the Audience

Who We Are

Erin Boon

Data Scientist
BLS
Office of Survey Methods Research

Panel Moderator

Ayme Tomson

Data Scientist
BLS
Office of Prices & Living Conditions

Lead Data Scientist for Consumer Price Index Housing Address Matching Project

Melissa Pollock

Economist
BLS
Office of Prices & Living Conditions

Product owner for Consumer Expenditure Diary Autocoder

Daniel Todd

Data Scientist
BLS
Office of Compensation and Working Conditions

Product owner for Survey of Occupational Injuries and Illnesses (SOII) Autocoder

Consumer Price Index Housing Address Matching

Ayme Tomson

Data Scientist

BLS Office of Prices & Living Conditions

Division of Consumer Prices and Price Indexes

Office of Prices and Living Conditions (OPLC)

Programs:

- Consumer Price Index (CPI)
- Producer Price Index (PPI)
- U.S. Import and Export Price Indexes (MXPI)
- Consumer Expenditures (CE)

OPLC Data Science Team

Support OPLC program economists and statisticians through state-of-the-art computing and statistical methods to ensure the accuracy, timeliness, and relevance of OPLC's outputs.

Projects:

- Alternative Data
- Data Imputation
- Data Analytics
- Automation
- Code Translation / Modernization
- Webscraping
- Dashboards / Visualization Tools
- API Development

OPLC Data Science Team

Support OPLC program economists and statisticians through state-of-the-art computing and statistical methods to ensure the accuracy, timeliness, and relevance of OPLC's outputs.

NLP Projects:

- Alternative Data
- Data Imputation
- Data Analytics
- Automation

Project Description & Goal

- The Consumer Price Index (CPI) Housing team verifies BLS subsidy and unit level data. This project compares matching techniques using the Housing and Urban Development (HUD) data as a supplementary data source of subsidy and unit level address information.

- Goal: Use one year of BLS and HUD data to assess current matching methodology and explore address matching improvements.

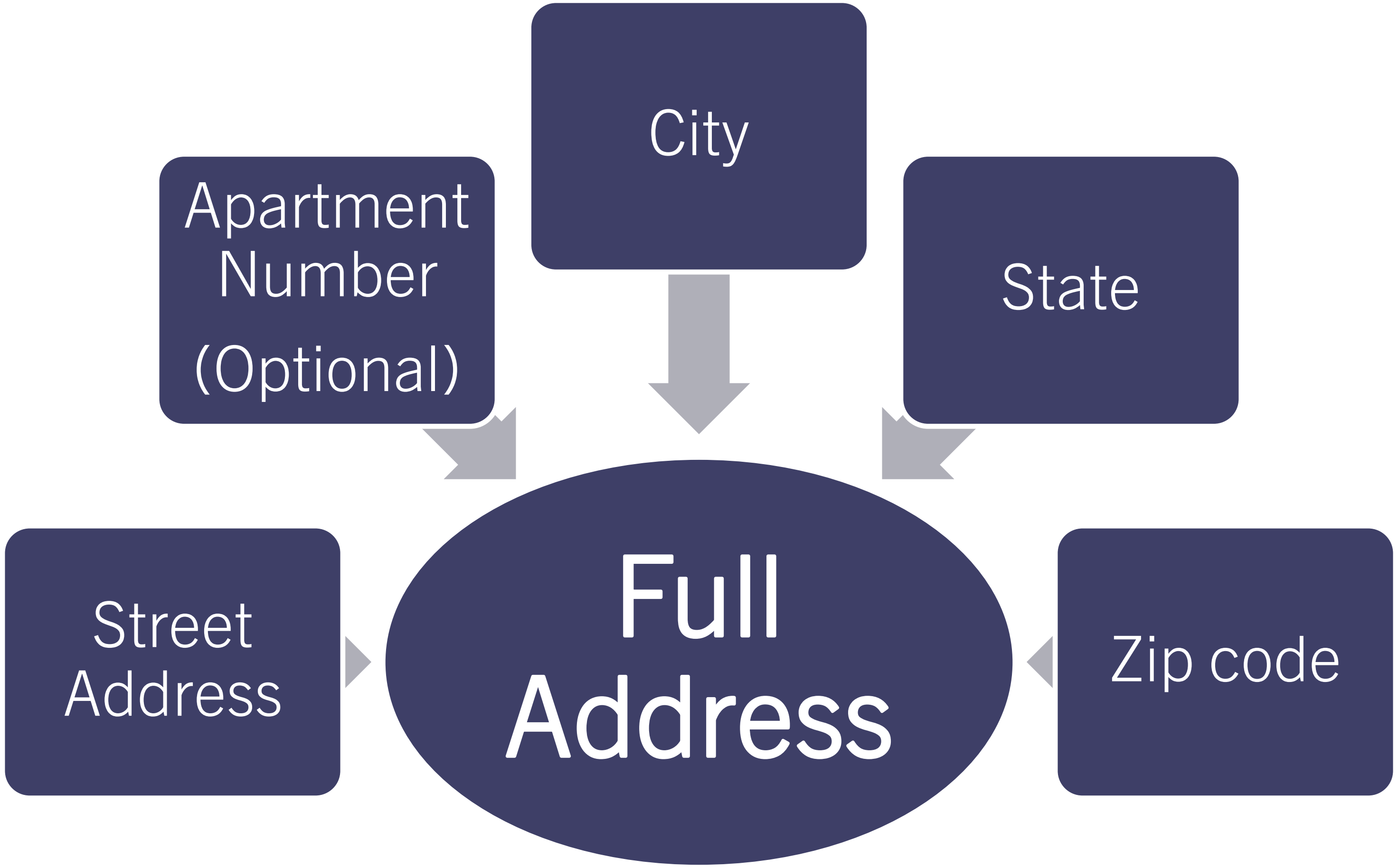
Project Members

- OPLC Data Scientists
  - Ayme Tomson
- Consumer Price Index (CPI) Housing Team
  - Brian Adams
  - Craig Brown
  - Austin Enderson-Ohrt
  - Ben Houck
  - Paul Liegey

Address Matching vs Geocoding

- Geocoding – estimating the physical location of an address
- Address Matching – determining if two addresses represent the same physical space

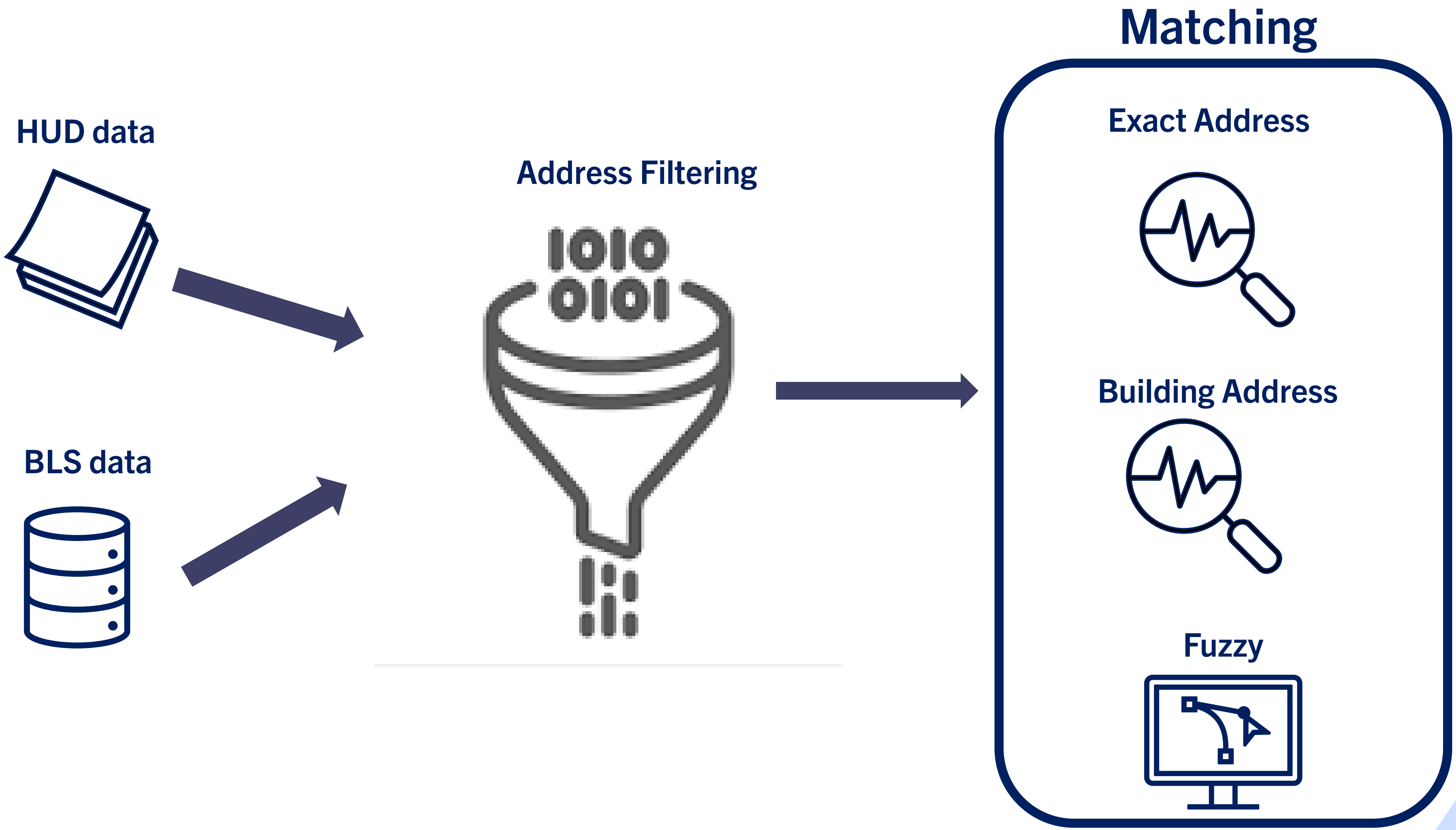
Address Matching



Address Matching Methodology

- Rule Based (Current approach)
  - Deterministic
  - Binary (match or non-match)
- Fuzzy Matching
  - Probabilistic
  - Continuous (probability of match)

Project Flow



Address Matching Challenges

Challenge	Example 1	Example 2
Null Values	NA	NaN
Missing Values / Imputation	Imputed from another column (zip, state)	Find correct data using data linkage (zip, state)
Punctuation	Dashes (–)	Numbers
Size	Small “ground truth” data set	Millions of a rows, tens of columns
Human Error / Variance	Competing Apt numbers	Regional language differences
Matching prioritization	State vs Zip vs Street address vs Apt number	Apt number exists but is missing
Matching Threshold	Exact Matching	Fuzzy Matching

Project Methodology

- Current literature that suggests a 90% matching is possible
  - Used linkage data not available to the CPI Housing Team
  - Used linkage data containing PII
  - More relaxed definition of “exact” address matching

Project Outcomes

- The Python approach replicated (could not significantly improve) the existing SAS approach already in use by the CPI Housing Team.

- The requirement for an exact address match limits the coverage available with the HUD data

Consumer Expenditure Diary Autocoder

Melissa Pollock

Economist

BLS Office of Prices & Living Conditions

Division of Consumer Expenditure Surveys

Consumer Expenditure Surveys Quick Facts

The surveys are the only federal government data collection effort to obtain information on the complete range of consumers' expenditures, income, and demographic characteristics, directly from consumers.

2 Surveys

The **Interview** survey collects detailed data on major and/or recurring expenditures for periods of 3 months or longer; the **Diary** survey collects records for smaller, more frequently purchased items. Both include household characteristics questions to record demographic information.

United States Census Bureau

National Processing Center (NPC)

Data are collected by Census for BLS: **Interview** expenditures via computer assisted personal interview (CAPI) instrument, **Diary** via respondent in a self-administered diary.

CE data are used by the **Consumer Price Index** to weight its price indexes, inform the study of population segments, and are inputs to other governmental agency statistics and private sector organizations.

The Diary Survey

ILLUSTRATIVE EXAMPLE

Clothing, Shoes, Jewelry, and Accessories

What did you buy or pay for?	Cost without tax		Was the item for: 1 Child Under 2	2 Boy 2-15	3 Girl 2-15	4 Man 16 & over	5 Woman 16 & over	Name of Store or Website where purchased
dress shirts	75	00					X	Dillards.com
running shoes	69	00					X	↓
wallet	29	00				X		
baseball cap	14	99		X				Target
bib	3	50	X					Sweet Dreams boutique
necklace	250	00					X	Olde Towne jewelry
non-prescription sunglasses	59	00					X	Walmart.com
~~child's costume (returned for refund)~~	~~15~~	~~00~~	~~X~~					~~Partysupply.com~~

Daily expenses are **recorded directly by the respondent** over two consecutive one-week periods

Four expenditure sections:

- Clothing, shoes, jewelry, accessories
- Food & drinks for home consumption
- Meals outside the home
- All other expenditures

Item Coding Example

sweatpants
bell-bottoms
PANTS
Work pants
corduroys
JEANS
joggers
Khakis
Linen trousers
cargo pants
High-waisted flared denim
DRESS SLACKS
Maternity pants
Yoga tights

All are grouped together in
SLACKS
Item code 410060

Item Coding Prior to 2024

WHAT	**30K Diary records** for each month of data manually keyed in and assigned an item code
HOW LONG	**8 weeks** of keying and coding for each month of Diary data
ACCURACY	**11.8%** of NPC-coded records are misclassifications

Opportunity

- **Reduce processing time**: speed up item code assignment
- **Uphold data accuracy**: reduce (or, at least, do not increase) the volume of item code misclassifications
- **Cost savings:** reduce the cost of Diary digitization

CE Diary Autocoder Overview

The CE Diary Autocoder is not a single model but, instead, a series of approaches to arrive at an item code assignment.

MEALS OUTSIDE THE HOME

ALL OTHER EXPENDITURE SECTIONS

RULES-BASED

Applies **if-then logic** to link the Diary entry to a corresponding item code

DIRECT MATCH

Looks for **identical text string matches** between the Diary entry and item code descriptions

MACHINE LEARNING

Predicts an item code using a **classification algorithm** learned from labeled historical data

{Used for additional validation against ML predictions}

SIMILARITY RATIO

Looks for **very similar text strings** between the Diary entry and item code descriptions

COMBO CODES

Looks for best possible matches for **multiple expenses within one Diary entry** *(e.g., "shirts and pants")*

An estimated 10-20% of this subset will have low probability of a correct prediction and be ***flagged for human review***

Method Details and Results

DIRECT MATCH

- Multi-step processing of the incoming Diary entry
- Compare against a maintained robust dictionary
- ~50% of records can be direct matched

MACHINE LEARNING

- Random Forest model built for each of the 4 record types using 2 years of training data
- Accuracy, Precision, Recall, and F1 used to evaluate model performance
- Low confidence predictions are flagged for human review

SIMILARITY RATIO

- For all records with an ML prediction, similarity ratio is calculated for the processed Diary entry and the predicted item code description
- *Formula:*
  $$\frac{2 * \text{number of matching characters}}{\text{Total number of characters}}$$

Accuracy same as Census coding

Estimated 73% reduction in processing time

SOII Autocoder

Daniel Todd

Data Scientist

BLS Office of Compensation and Working Conditions

Compensation Research and Program Development Group

Survey of Occupational Injuries and Illnesses (SOII)

- Establishment survey
- >200k injuries reported/year
- Information such as:
    - Job title
    - Source of injury
    - Part injured
    - Etc.

SOII Case Coding Example

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

Occupation: 37-2011 (Janitor)

Nature: 111 (Fracture)

Part: 420 (Arm)

Event: 422 (Fall, slipping)

Source: 6620 (Floor)

Why Computer-Assisted Coding/Autocoder?

It started more than 12 years ago when all codes were assigned manually.

Manual coding was time and resource intensive.

People weren't coding consistently across regions.

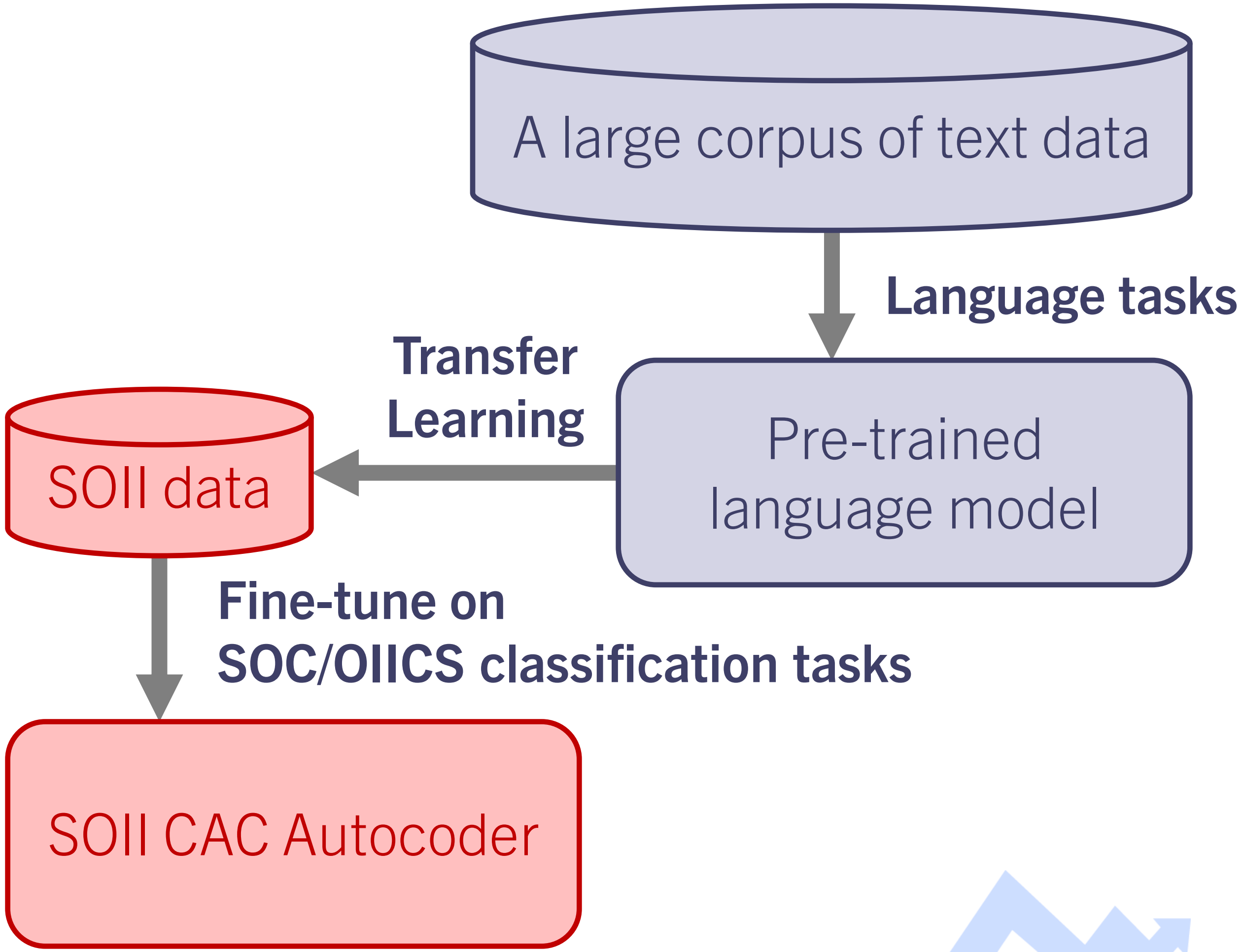
- Two experts coding exact same narratives ~70% agreement

Can computers help?

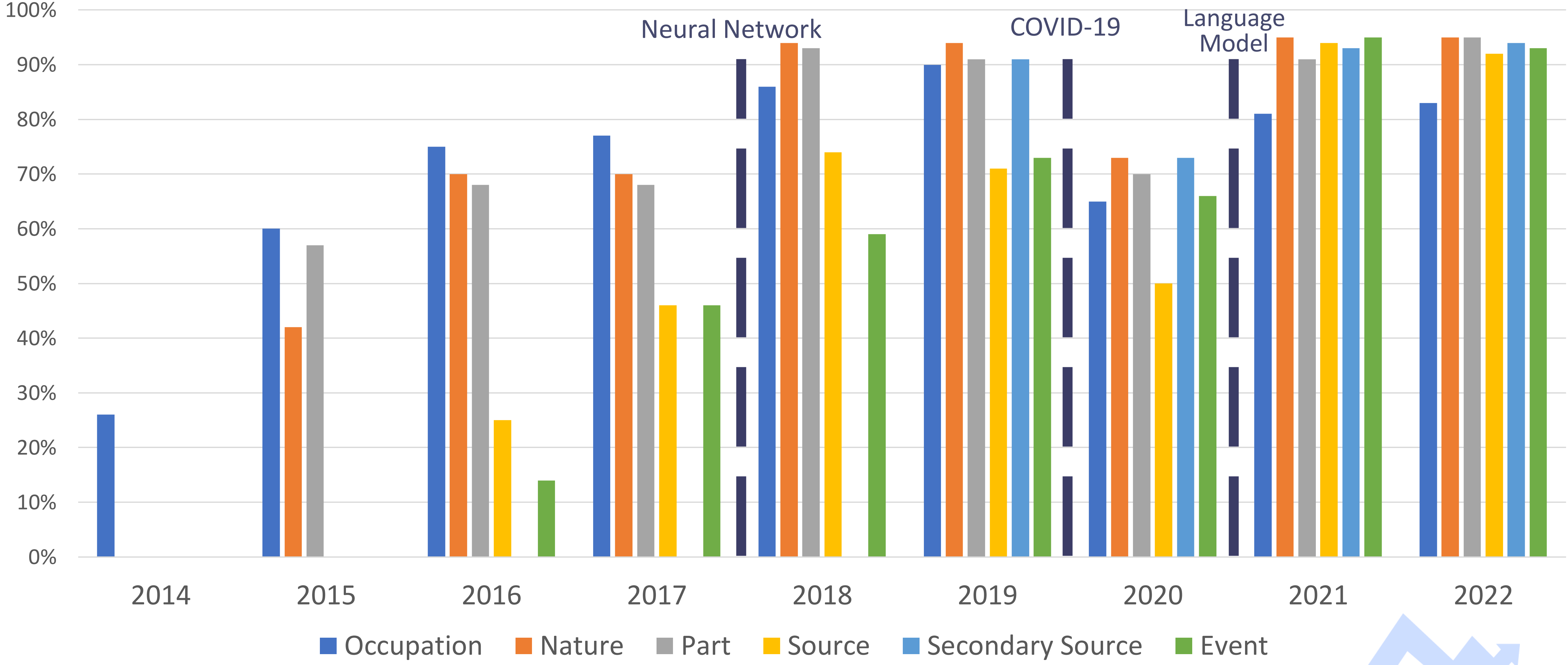
Current SOII Autocoder: Language Model

Timeline:

- 2014: Logistic Regression/Bag of Words
- 2018-2020: LSTM model
- 2021: Transfer learning using transformer model (like Chat-GPT)



Percent of SOII codes automatically assigned by survey year

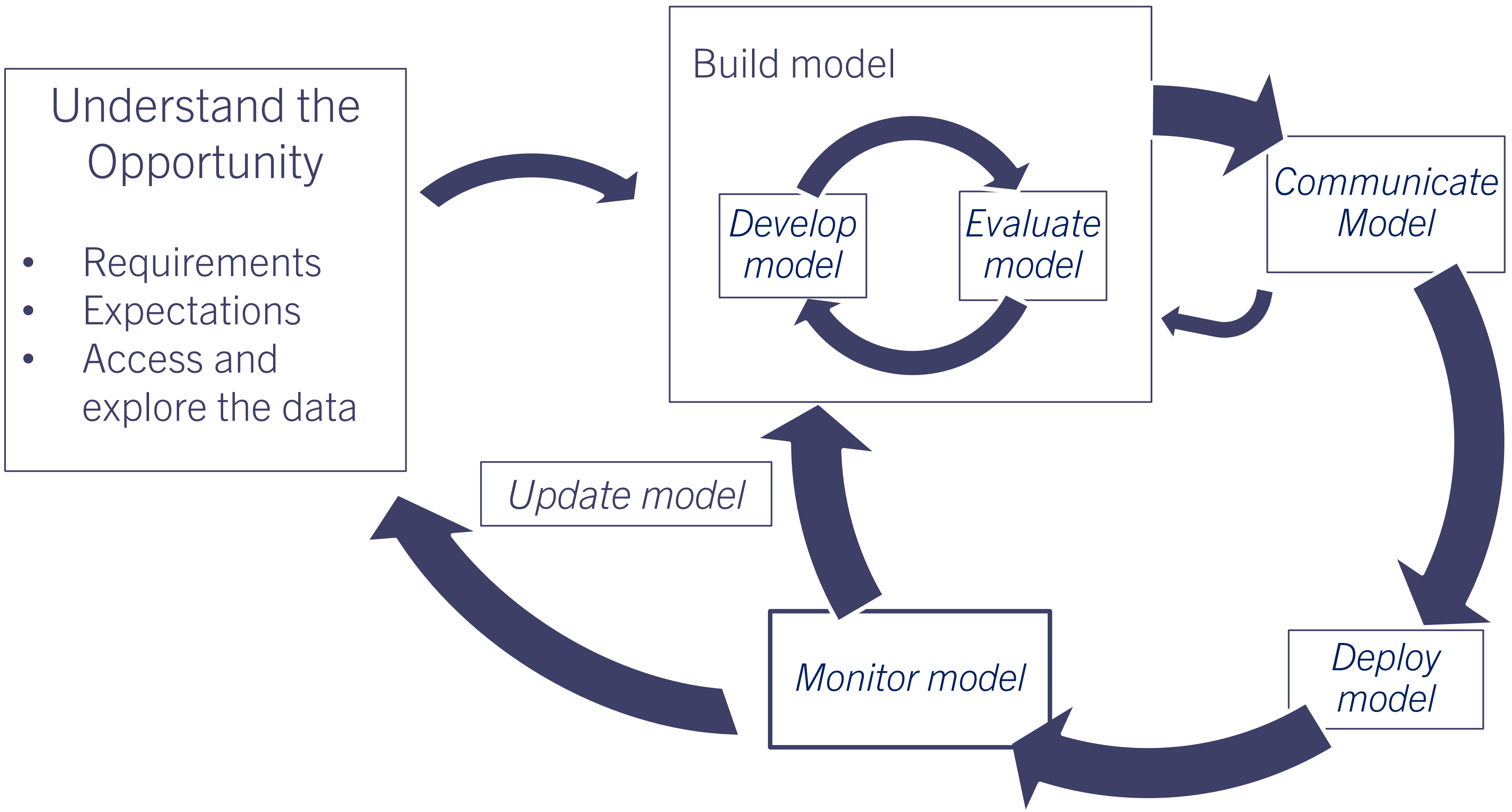


Autocoding Process Safeguards

- **Human in the loop**: All predictions made are reviewed by human staff.
- **Performance measurement using gold standard data**: Prior to deployment, models are evaluated against gold standard dataset labelled by subject matter experts

Q&A

Model as Product: Phases of Development



Contact Information

Erin Boon

Data Scientist
BLS
Office of Survey Methods Research
Data Science Research Center

Boon.Erin@bls.gov

Ayme Tomson

Data Scientist
BLS
Office of Prices & Living Conditions
Division of Consumer Prices and Price Indexes

Tomson.Ayme@bls.gov

Melissa Pollock

Economist
BLS
Office of Prices & Living Conditions
Division of Consumer Expenditure Surveys

Pollock.Melissa@bls.gov

Daniel Todd

Data Scientist
BLS
Office of Compensation and Working Conditions
Compensation Research and Program Development Group

Todd.Daniel@bls.gov