Current Scope of AI Implementations at BLS

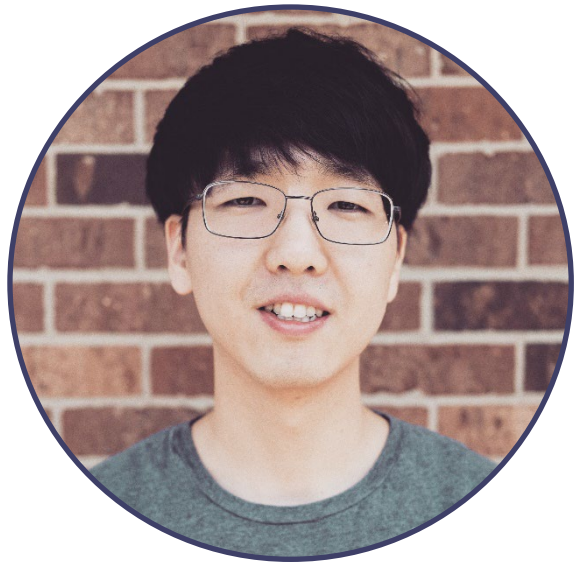
David H. Oh

Supervisory Data Scientist

AI Day for Federal Statistics: CNSTAT Public Event

May 02, 2024

About Me



David H. Oh

- Supervisory Data Scientist
- Oversees various data science operations in the Office of Compensation and Working Conditions
- Joined the Bureau of Labor Statistics in 2017 as an Economist and transitioned into a Data Scientist role in 2020

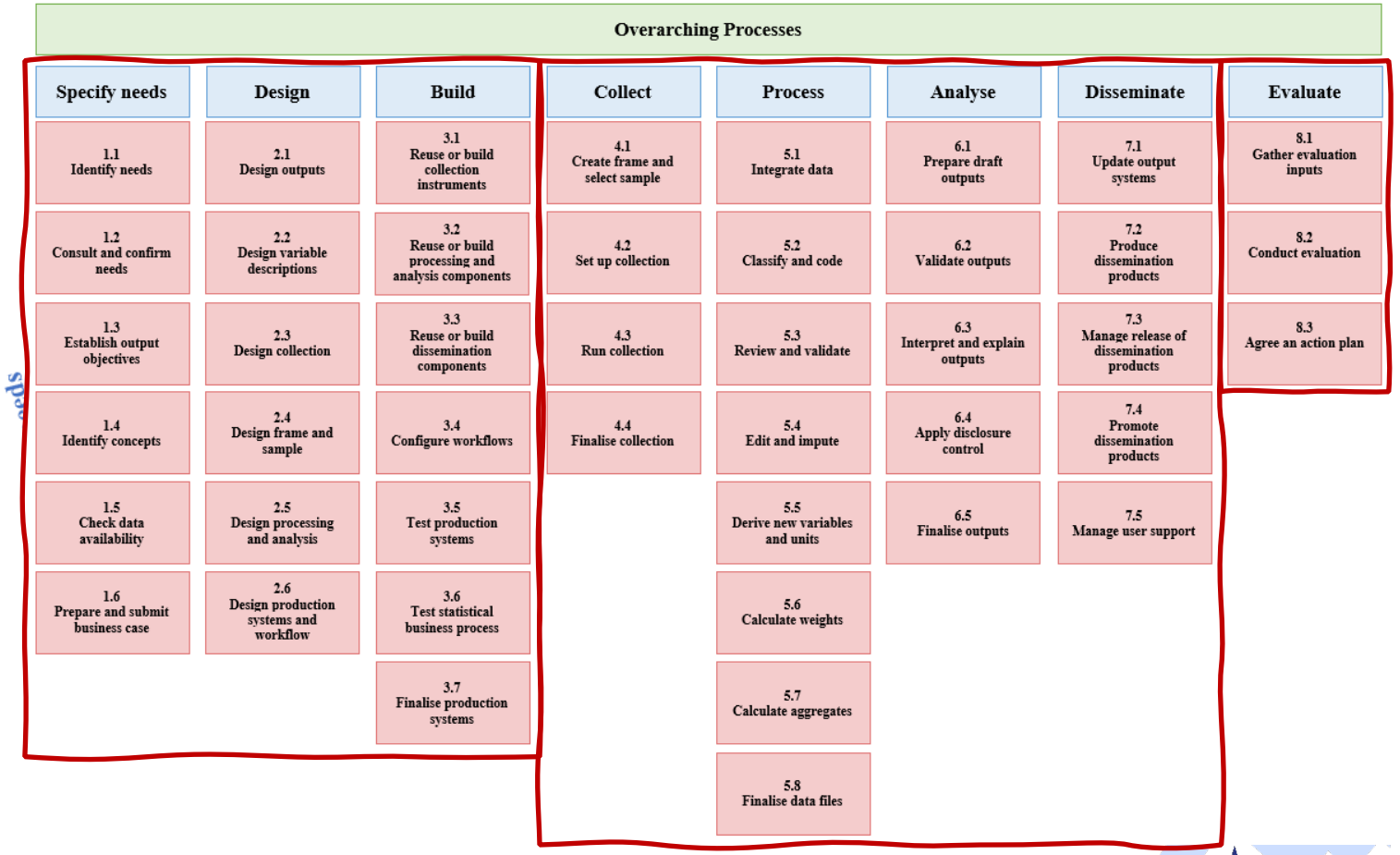
Overview

- Discuss framework for determining the scope of AI implementations
- Provide an overview of current AI implementations at BLS
- Discuss future AI opportunities

General Statistical Business Process Model

- The General Statistical Business Process Model (GSBPM) describes and defines the set of business processes needed to produce official statistics
- First developed in 2008 by the Joint UNECE/Eurostat/OECD Group on Statistical Metadata (METIS)
- Three levels
  - Level 0: the statistical business process
  - Level 1: the eight phases of the statistical business process
  - Level 2: the sub-processes within each phase
- https://statswiki.unece.org/display/GSBPM

General Statistical Business Process Model



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

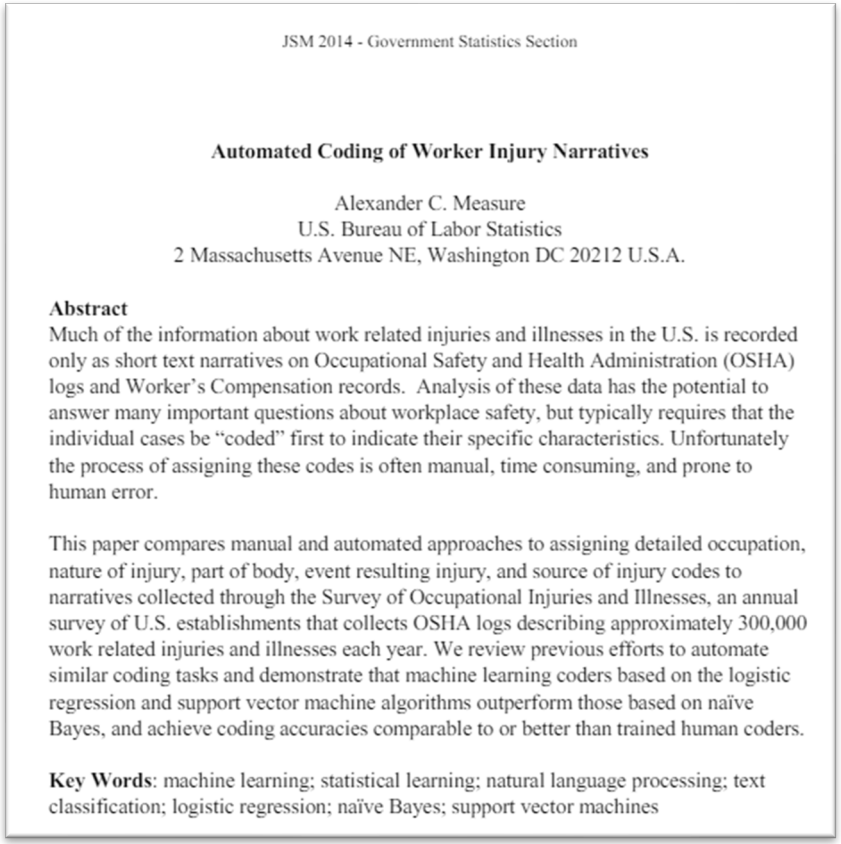
Labor market added 216,000 jobs in December, capping year of big gains

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

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

AI Implementations at BLS

- Developed by Data Scientists, Economists, and Statisticians in the program offices to improve specific operational needs
- Developed in house using open-source languages like Python and R
- Involves leveraging text data (i.e., natural language, addresses, etc.)
- Over a decade years of experience experimenting with AI implementation



JSM 2014 - Government Statistics Section

Automated Coding of Worker Injury Narratives

Alexander C. Measure
U.S. Bureau of Labor Statistics
2 Massachusetts Avenue NE, Washington DC 20212 U.S.A.

Abstract
Much of the information about work related injuries and illnesses in the U.S. is recorded only as short text narratives on Occupational Safety and Health Administration (OSHA) logs and Worker's Compensation records. Analysis of these data has the potential to answer many important questions about workplace safety, but typically requires that the individual cases be "coded" first to indicate their specific characteristics. Unfortunately the process of assigning these codes is often manual, time consuming, and prone to human error.

This paper compares manual and automated approaches to assigning detailed occupation, nature of injury, part of body, event resulting injury, and source of injury codes to narratives collected through the Survey of Occupational Injuries and Illnesses, an annual survey of U.S. establishments that collects OSHA logs describing approximately 300,000 work related injuries and illnesses each year. We review previous efforts to automate similar coding tasks and demonstrate that machine learning coders based on the logistic regression and support vector machine algorithms outperform those based on naïve Bayes, and achieve coding accuracies comparable to or better than trained human coders.

Key Words: machine learning; statistical learning; natural language processing; text classification; logistic regression; naïve Bayes; support vector machines

AI Implementations at BLS

Automated Coding

Document Processing

Record Matching

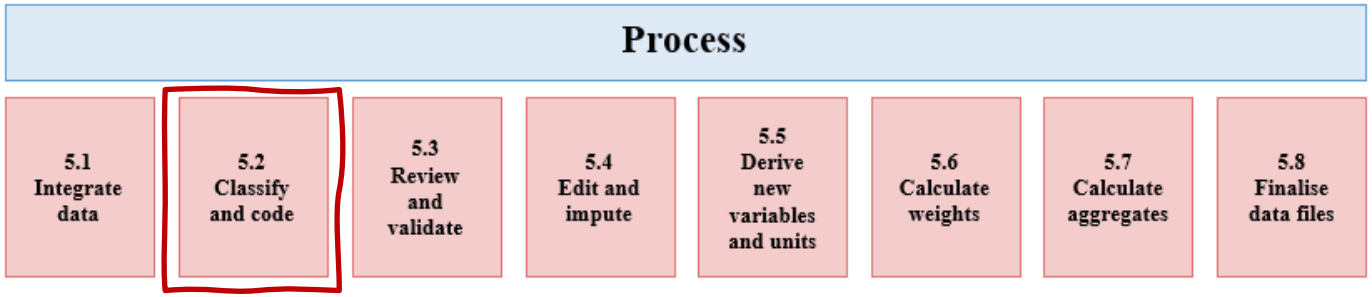
Data Imputing

Automated Coding at BLS

Automated Coding

Using machine learning to automatically classify open-ended text statements about job titles, workers injuries, or products and services into SOC, OIICS, and item coding systems respectively

- Supervised machine learning, leveraging large sets of previously labeled data
- Complexity of the ML models ranges from regularized logistic regressions to transformers

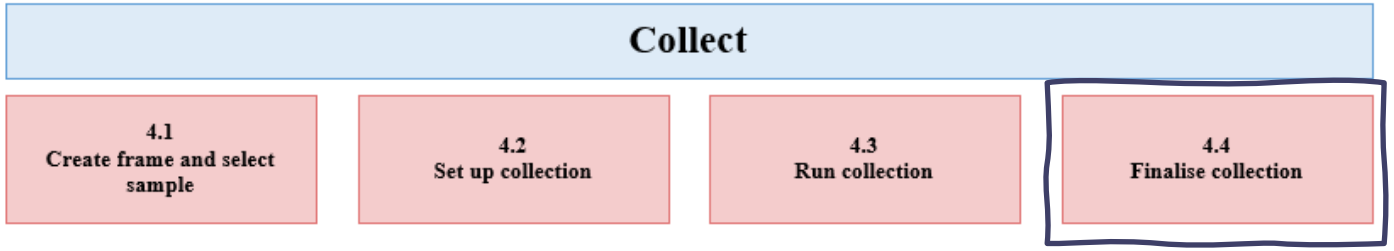


Document Processing at BLS

Document Processing

Using language models to automatically process documents like job descriptions, news articles, or summary of benefit coverages to extract relevant data elements and/or to summarize their content

- Documents from respondents or other sources contain useful information
- Language models offer a variety of functions, such as named-entity recognition, summarization, and question-answering

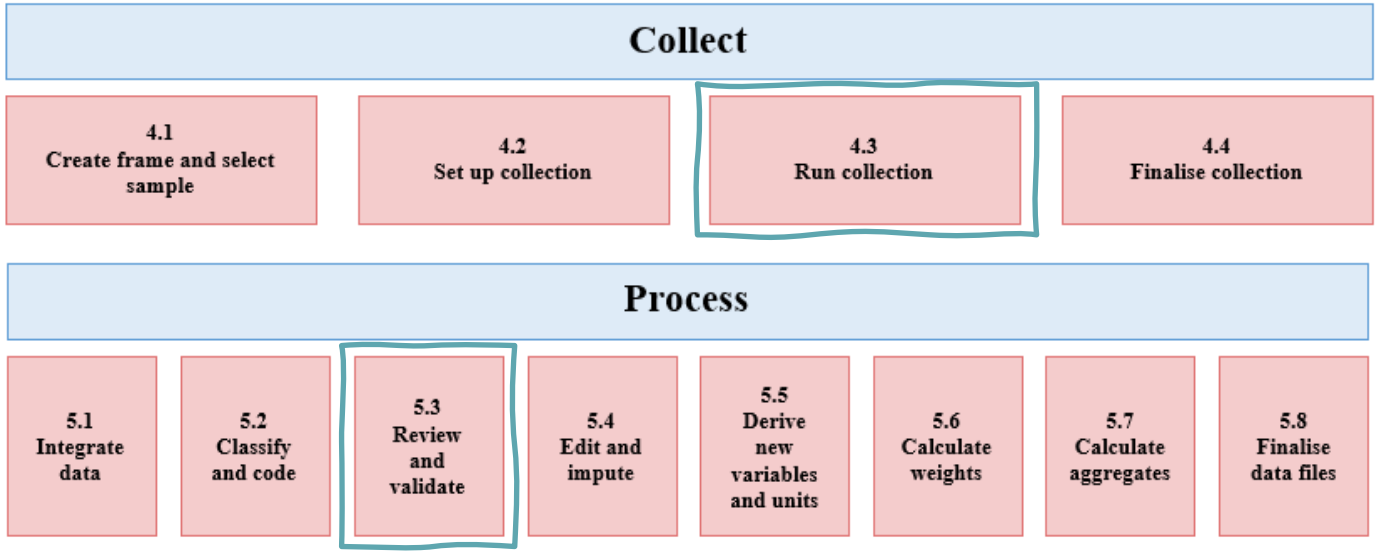


Record Matching at BLS

Record Matching

Leveraging word embeddings to measure distance between two records for use in sample refinement as well as identifying missing or inconsistent data

- Variety of use cases: sample refinement, data review, exploration of new outputs

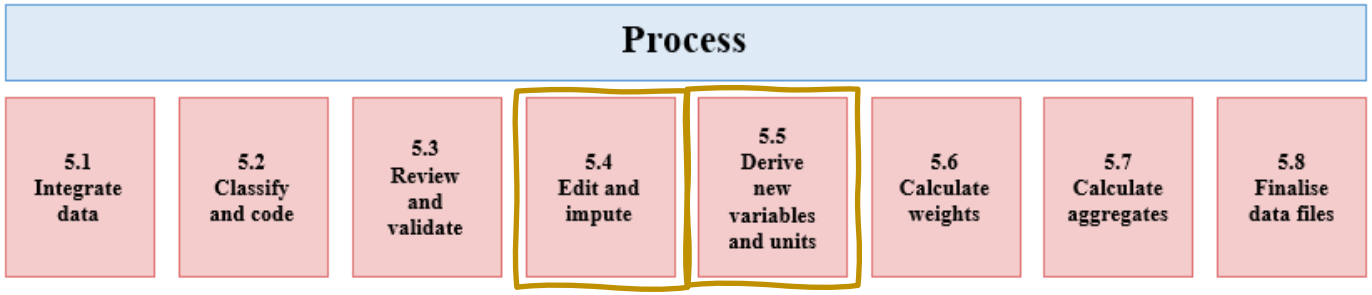


Data Imputing at BLS

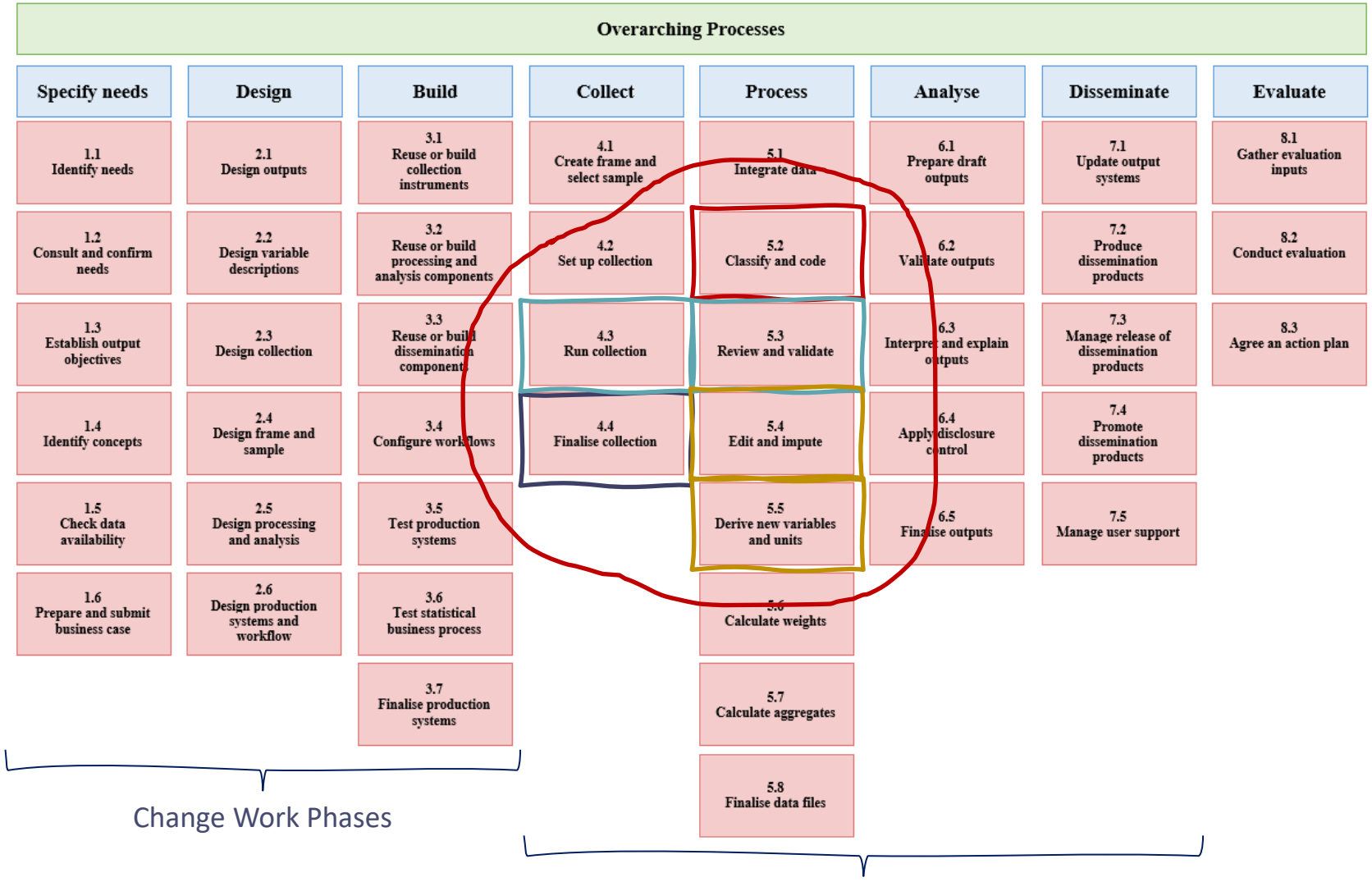
Data Imputing

Using machine learning to predict or estimate missing or new variables for use in producing official statistics, such as the measures of hours worked

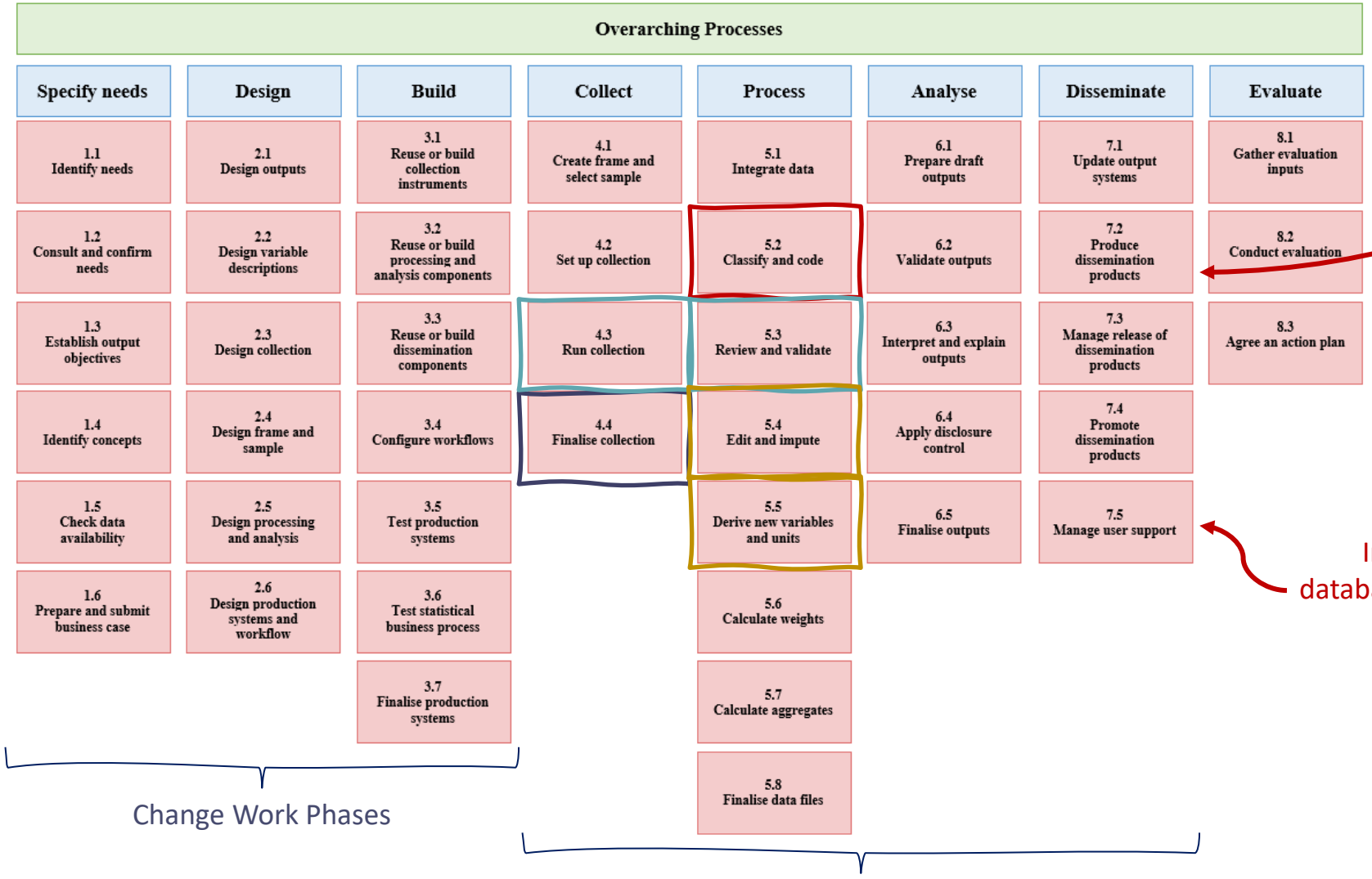
- Use of machine learning techniques over more traditional statistical approaches
- Supervised machine learning, leveraging already labeled data



General Statistical Business Process Model



Future AI Opportunities



Contact Information

David H. Oh

Supervisory Data Scientist

Compensation Research and Program Development Group

Office of Compensation and Working Conditions

Oh.David@bls.gov