Expanding Data Science at the U.S. Energy Information Administration

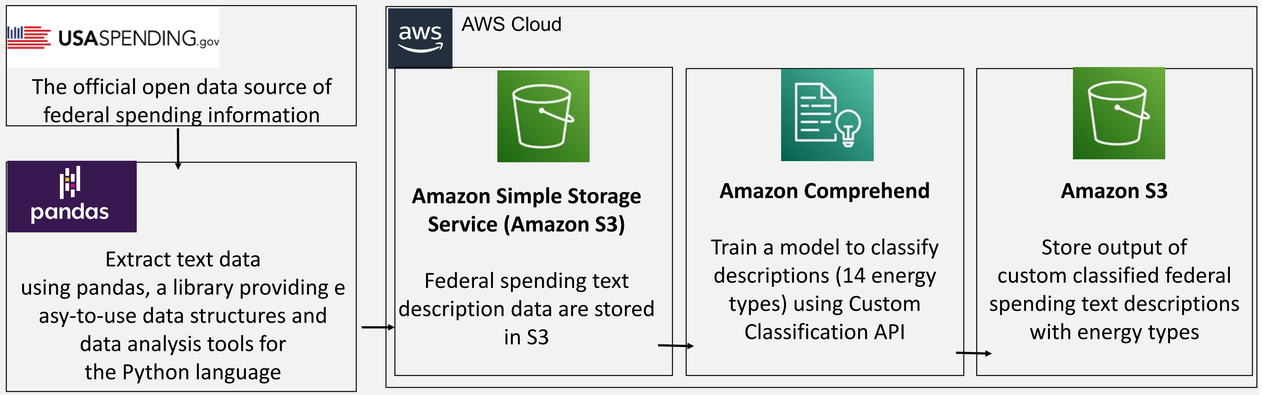
Katie Lewis (katie.lewis@eia.gov), Mark Schipper (mark.schipper@eia.gov) | U.S. Energy Information Administration (EIA)

Project highlight: Use of Natural Language Processing in the Subsidies Report

Background: EIA conducts a comprehensive data collection and analytic program that covers the full spectrum of energy sources, end uses, and energy flows. We also produce analytic reports when requested by Congress. One such Congressionally requested report is the *Federal Financial Interventions and Subsidies in Energy*, (the "Subsidies Report") which estimates how federal financial actions are distributed among a defined set of 14 energy types that make up the U.S. energy system. We developed a model that takes transaction descriptions as inputs and attempts to predict the corresponding energy subsidy category.

Goal: Classify nearly 80 million discrete federal budgetary award transactions across seven fiscal years.

Solution: Used AWS (Amazon) Comprehend, a natural language processing (NLP) service that uses machine learning to find insights and relationships in text, to classify individual direct and R&D expenditures downloaded from USASpending.gov. Classification is a two-step process: (1) train a custom classification model, (2) use model to classify data sources, which are subsequently reviewed by subject-matter experts.



Impact: We saved time with minimal costs, focusing subject-matter experts on reviewing near-final classification data, rather than ad-hoc work. AWS Comprehend improves the pre-processing pipeline by integrating named entity recognition with subject-matter experts.

Using data science to improve data quality and processing

- Identified misclassifications in administrative trade data using product descriptions
- Recoded other specify write-in responses using NLP
- Used fuzzy matching to compare addresses for frame maintenance
- Used satellite data to create frame for survey of buildings
- Identified potential misclassifications in survey responses using NLP
- Used lasso regression to model response propensities
- Used web scraping to collect administrative/third party data

Training and education

Basic Python programming → Data wrangling → Text analysis → Statistical modeling of data → Causal inferences from statistical models → Visualization and storytelling

- Data science skills development pilot program
- Python training cohorts with data science modules
- Self-development with LinkedIn Learning, Percipio, webinars, & conferences
- Knowledge sharing with brown bags, guest speakers, data science handbook