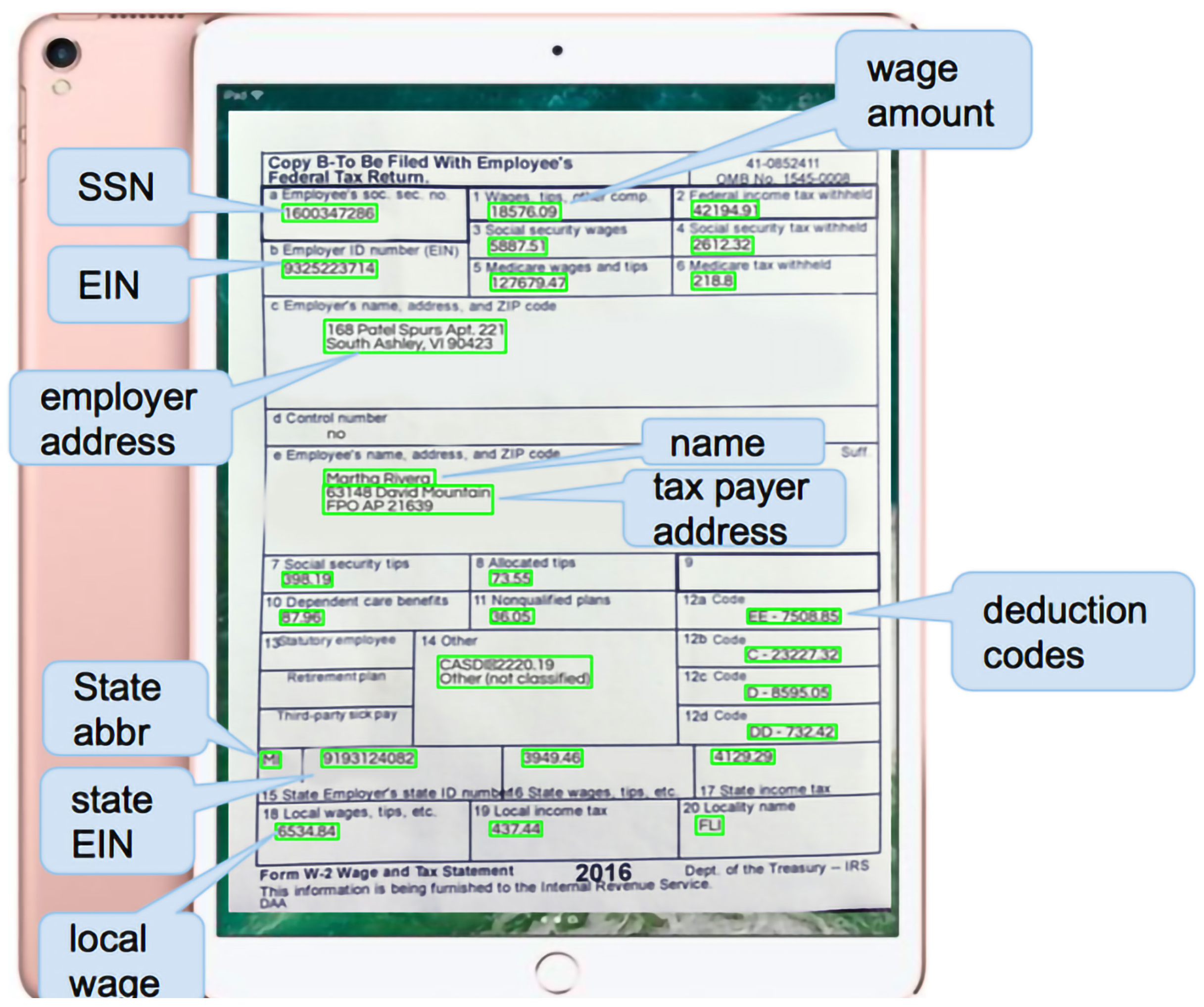
# Learning Information Extraction from Images of Structured Documents Using Synthetic Data and Conditional Random Fields (CRFs)

**Tharathorn (Joy) Rimchala**
*Intuit*

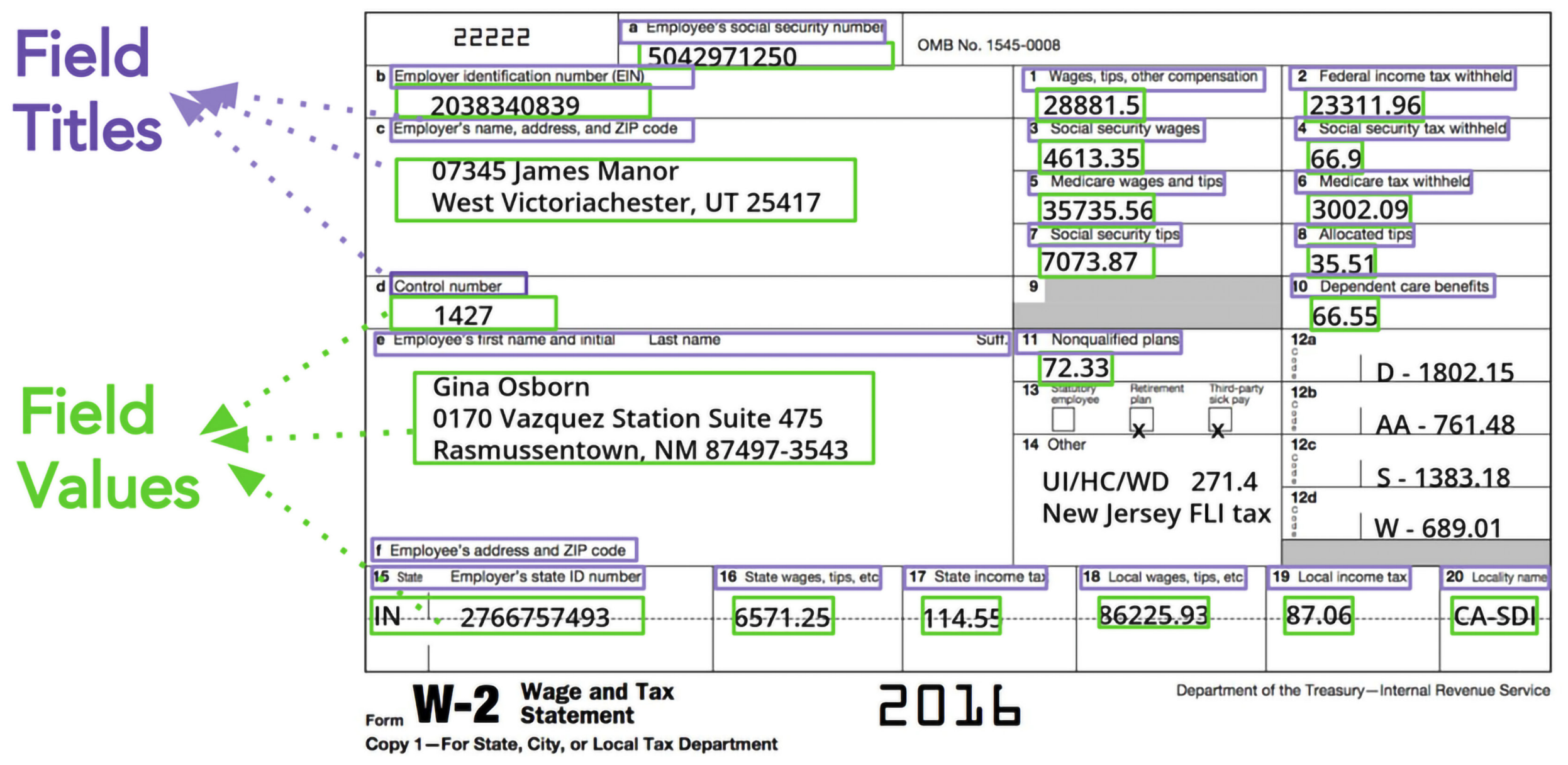
## Problem

### Why automatic information extraction for forms?



- Automatic information extraction (IE) from document or form images eliminates error-prone manual data entry so users don't have to do it.
- Existing machine learning-based techniques for IE rely on **expensive-to-acquire** manually annotated labeled data.
- Intuit solves this problem by building a data-driven synthetic data generation pipeline and by using a modified conditional random field (CRF) model for field extraction.

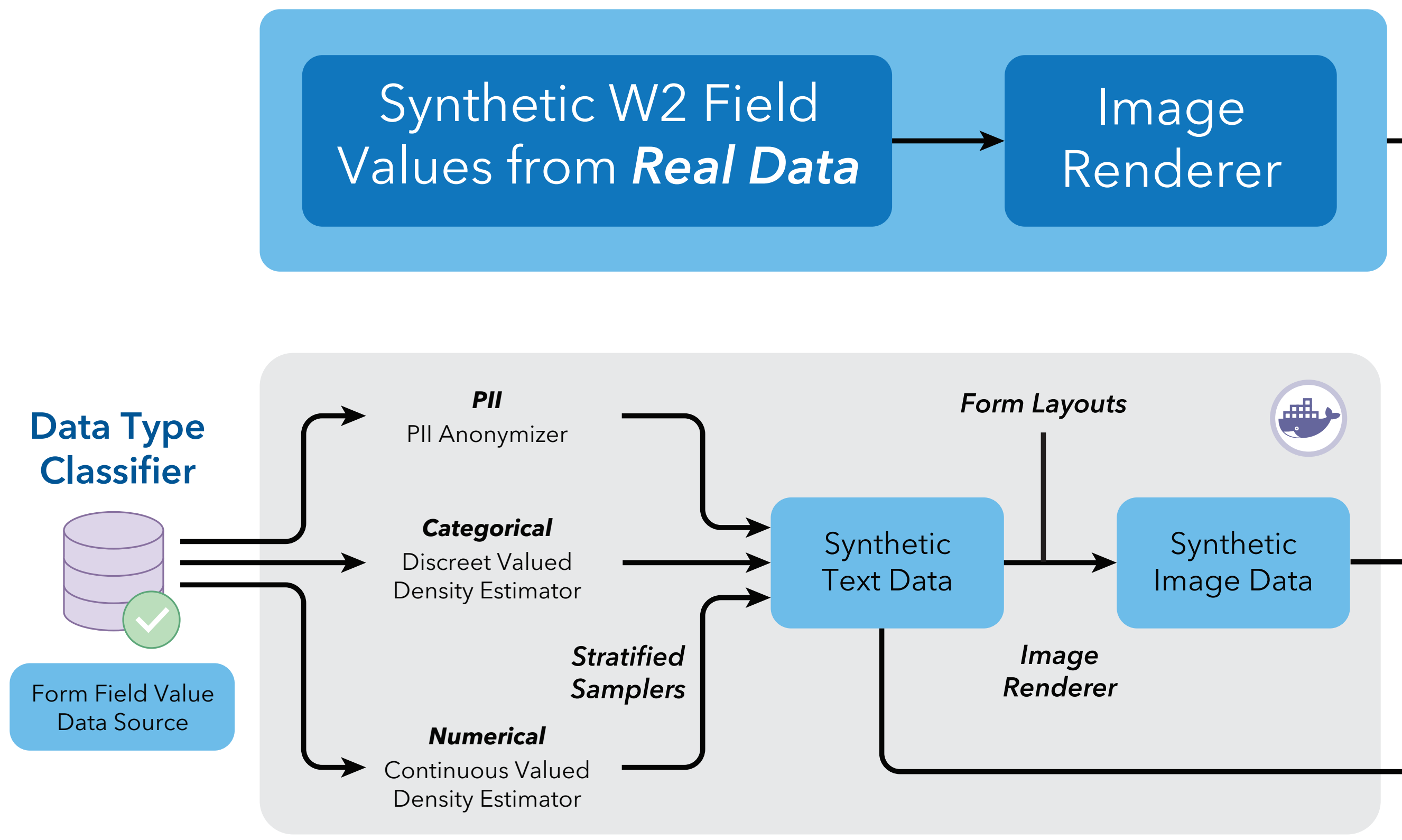
### Form automatic information extraction as a named entity recognition (NER) problem


Field Titles
Field Values

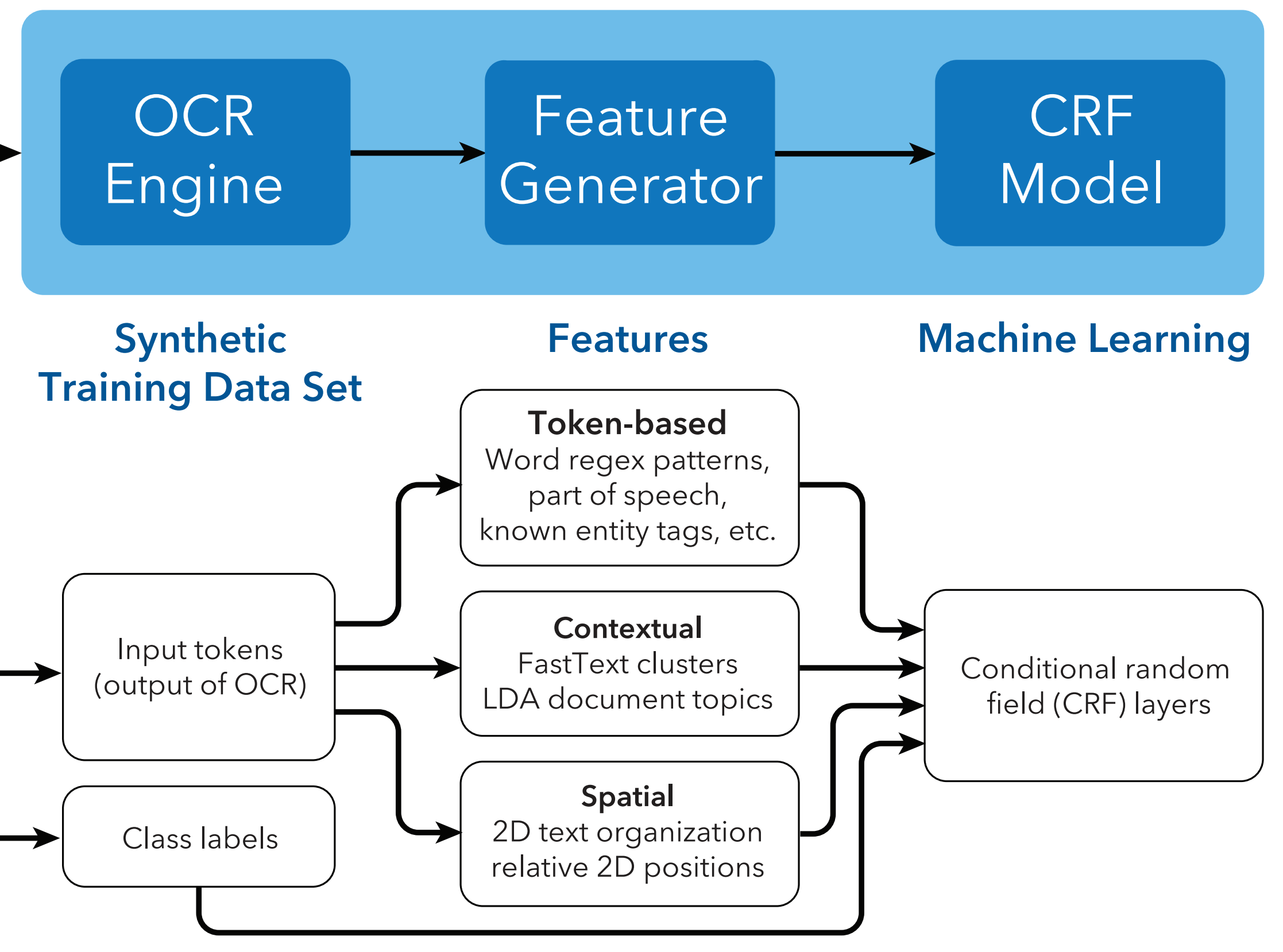
- Two main types of entities: form field titles and form field values.
- W2 form contains 32-35 fields, corresponding to ~70 entity classes.

## **Solution:** An End-to-End Learning Framework for Form Information Extraction

### Synthetic Data Generation



### Form CRF Model Training



#### The Pipeline Stages

- The generators learn to generate three main types of data distributions from millions of anonymized real electronic form field data.
- The synthetic data is rendered on variations of form with font variations.
- The entire pipeline is packaged into a single **ready-to-deploy** docker image.

#### Model Training

- 96K images, 48 variations, ~20 font variations and text localization
- 80/10/10 training/validation/test split
- L-BFGS with L2 regularization

## Results

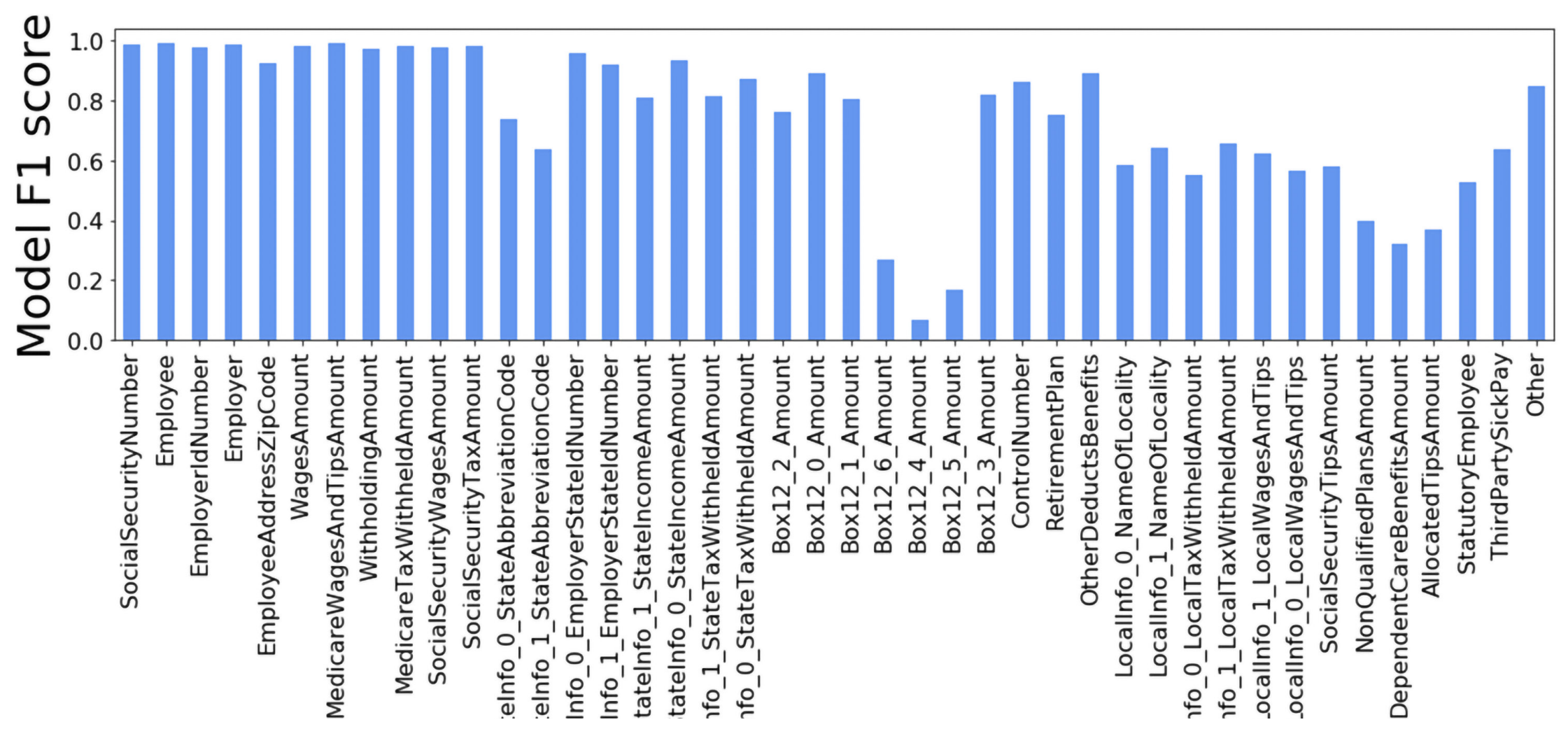
**Best NER-CRF Model Performance**
(aggregated across all classes)

| Metric | Field Title classes | Field value classes | High Usage* Field Value classes |
|---|---|---|---|
| Precision | 99.10% | 81.60% | 97.74% |
| Recall | 98.50% | 82.40% | 96.65% |
| F1 | 98.50% | 82.40% | 97.65% |

* High usage classes: classes that appear in more than 70% of all real W2 form data.

**NER-CRF Model Confidence Among Highly Used Classes**

| Field Class (entity we want to extract) | Model Confidence | Usage Rate |
|---|---|---|
| Employee Name | 99.19% | 99.86% |
| Employee Address | 92.31% | 99.82% |
| Employer Id Number (EIN) | 97.70% | 99.85% |
| Medicare Tax Withheld Amount | 98.18% | 98.39% |
| Medicare Wages And Tips Amount | 99.24% | 98.62% |
| Social Security Number (SSN) | 98.64% | 99.86% |
| Social Security Tax Amount | 98.23% | 97.23% |
| Social Security Wages Amount | 97.86% | 97.44% |
| Employer State Id Number (State EIN) | 95.93% | 87.34% |
| State Income Amount | 93.48% | 87.33% |
| Wages Amount | 98.25% | 99.77% |
| Withholding Amount | 97.12% | 98.42% |



### Model Performance:

- Model performance varies with the usage rate of field class in W2 Form.
- The best model yield **97.44% F1** score on classes that are **highly used field value class.**

### Next Steps:

- Deployment of the best NER-CRF.
- Exploring over-sampling to improve performance on classes that are not highly used.
- Exploring more powerful model: bidirectional Long Short Term Memory Conditional Random Field (biLSTM-CRF) using biLSTM as a feature extractor.