



# Multi-class Multi-label Patent Classification with Hybrid Neural Networks

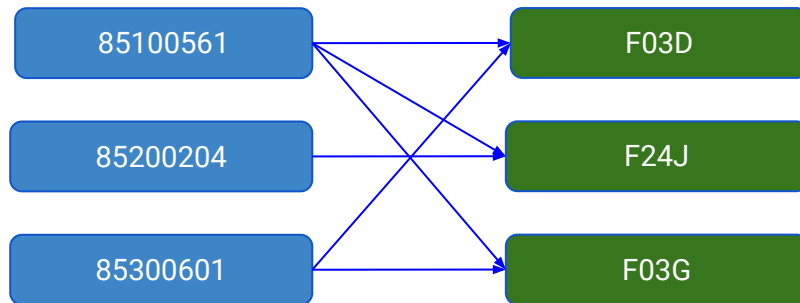
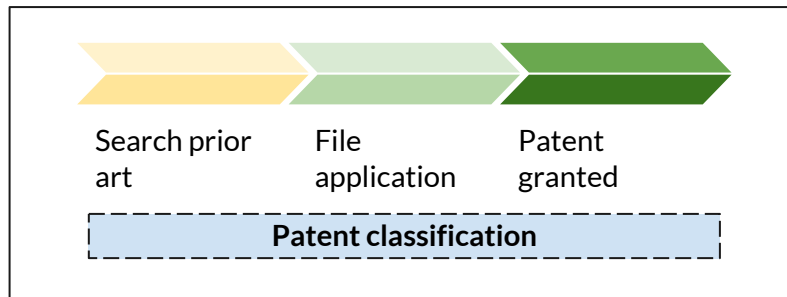
Saurav Datta, Arvinth Ganesan, Christina Papadimitriou

W266, Summer 2018

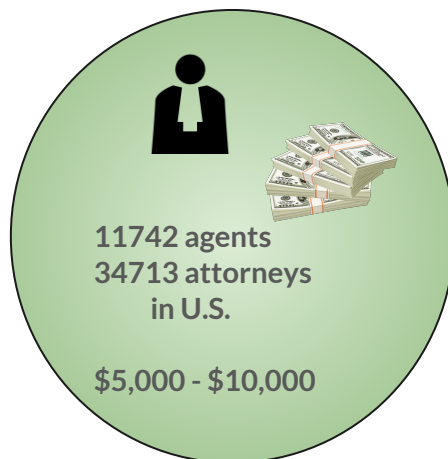
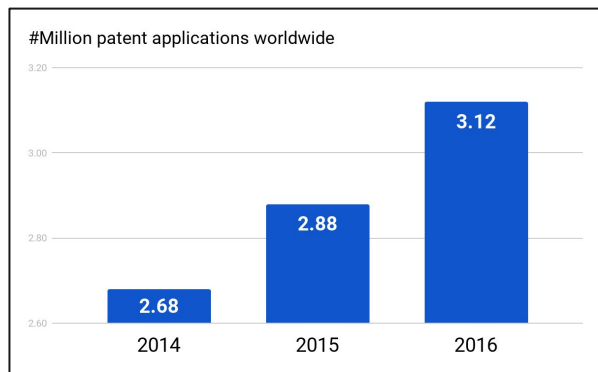
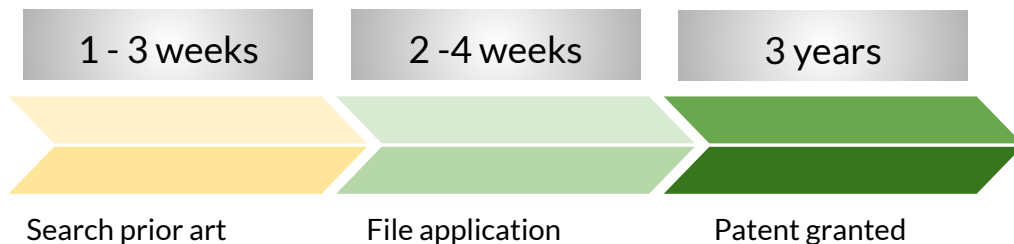
# Introduction

## Patent classification

- Integral part of patent application
- Makes it possible to search quickly for documents about earlier disclosures similar to or related to the invention for which a patent is applied for
- A patent can have multiple IPC labels
- PAC - patent automatic classification



# Challenges In Existing Process and PAC

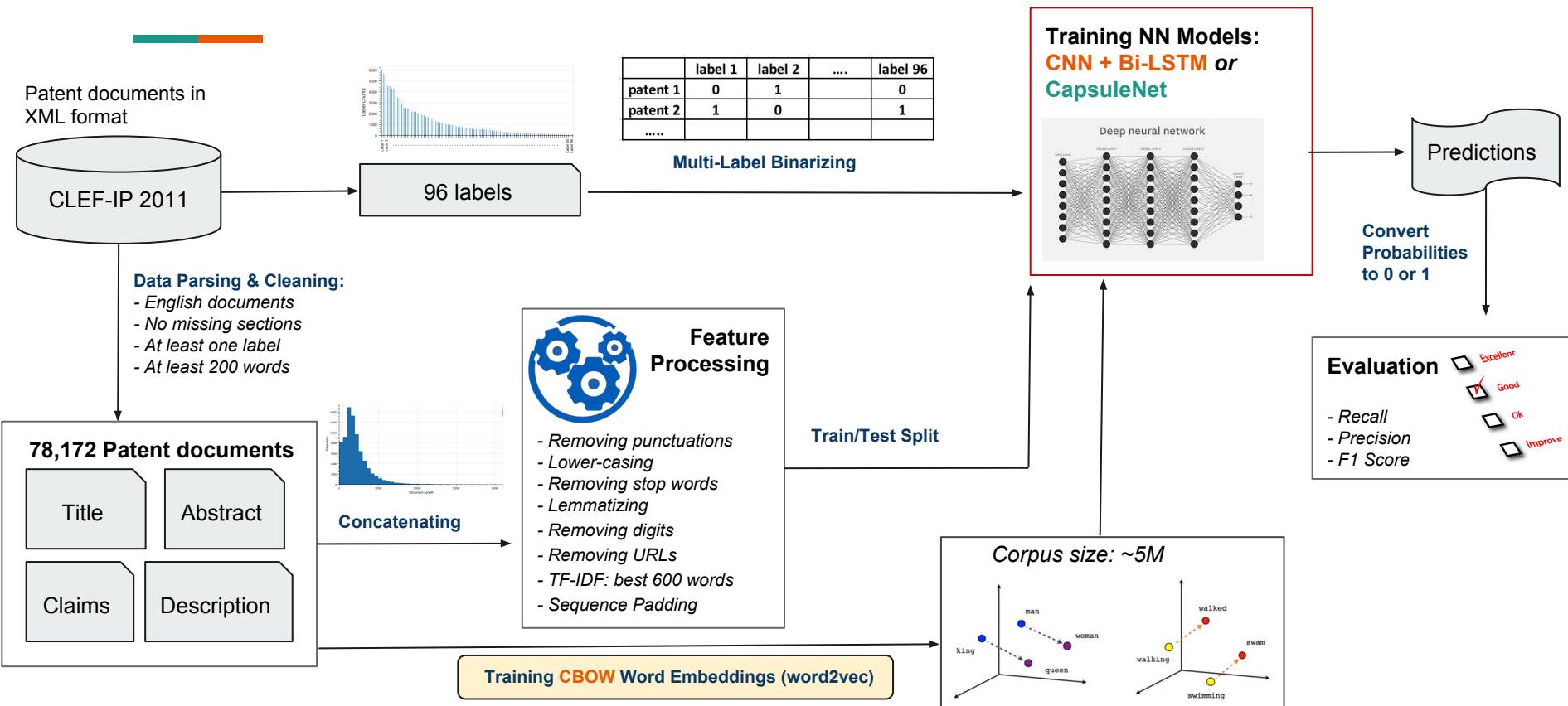


MultiLabel-  
MultiClass  
classification

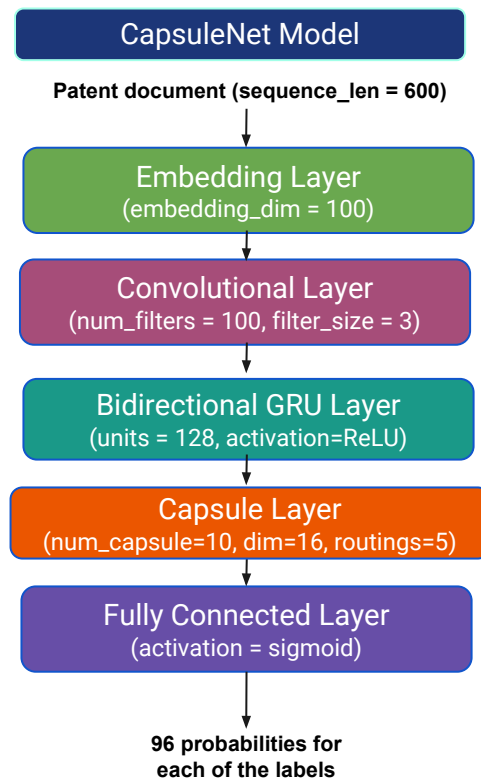
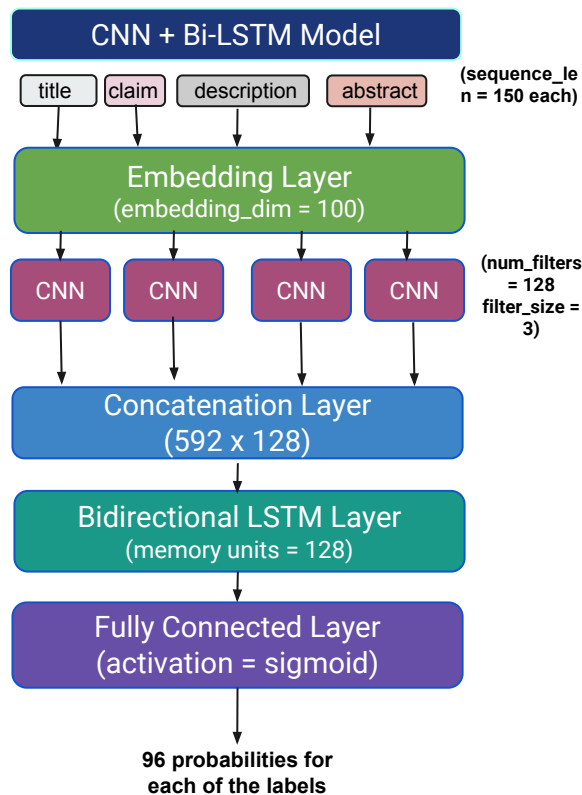
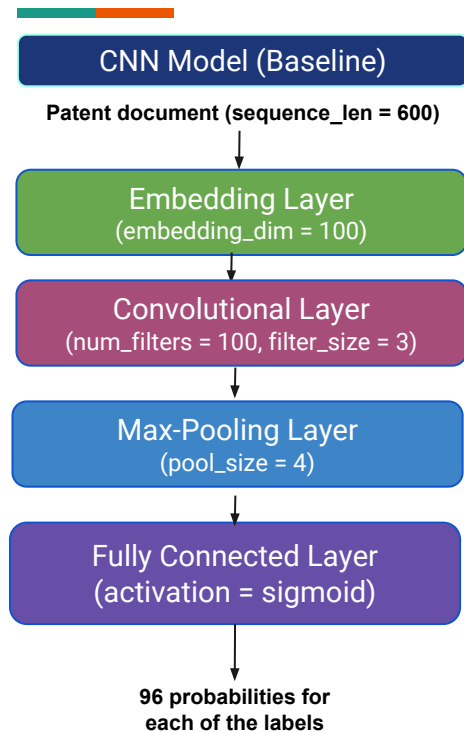
Non-standar  
d retrieval  
methods

Dependenc  
y on domain  
experts

# Process Flow



# Model Architectures



# Evaluation Metrics

**Precision:** the ratio of how much of the predicted is correct

**Recall:** the ratio of how many of the actual labels were predicted

We evaluate the below metrics when predicting the **top 1**, **top 2** and **top 10** IPC labels

## Document level calculations:

$$Precision = \frac{TP}{TP + FP} = \frac{trueLabels \cap predictionLabels}{predictionLabels}$$

$$Recall = \frac{TP}{TP + FN} = \frac{trueLabels \cap predictionLabels}{trueLabels}$$



## Aggregate calculations:

$$Precision_{total} = \frac{1}{TotalSamples} \sum_{n=i}^{TotalSamples} Precision_i$$

$$Recall_{total} = \frac{1}{TotalSamples} \sum_{n=i}^{TotalSamples} Recall_i$$

$$F1_{total} = 2 * \frac{Precision_{total} * Recall_{total}}{Precision_{total} + Recall_{total}}$$

# Results

	Top 1 Label			Top 2 Labels			Top 10 Labels		
	Precision	Recall	F1-Score	Precision	Recall	F1-Score	Precision	Recall	F1-Score
CNN (Baseline)	17.0%	13.2%	14.2%	11.3%	17.2%	13.1%	6.4%	48.5%	11.1%
CNN + Bi-LSTM	71.0%	59.2%	64.9%	47.7%	73.7%	57.9%	12.8%	93.0%	22.5%
Capsule Network	79.2%	65.5%	71.7%	53.1%	82.0%	64.5%	13.4%	97.3%	23.6%

- CapsuleNet overperforms all models including the baseline paper:
  - F1-score for top 1 label 71.7% (our CapsNet) vs 63.97% (CNN+Bi-LSTM from paper)