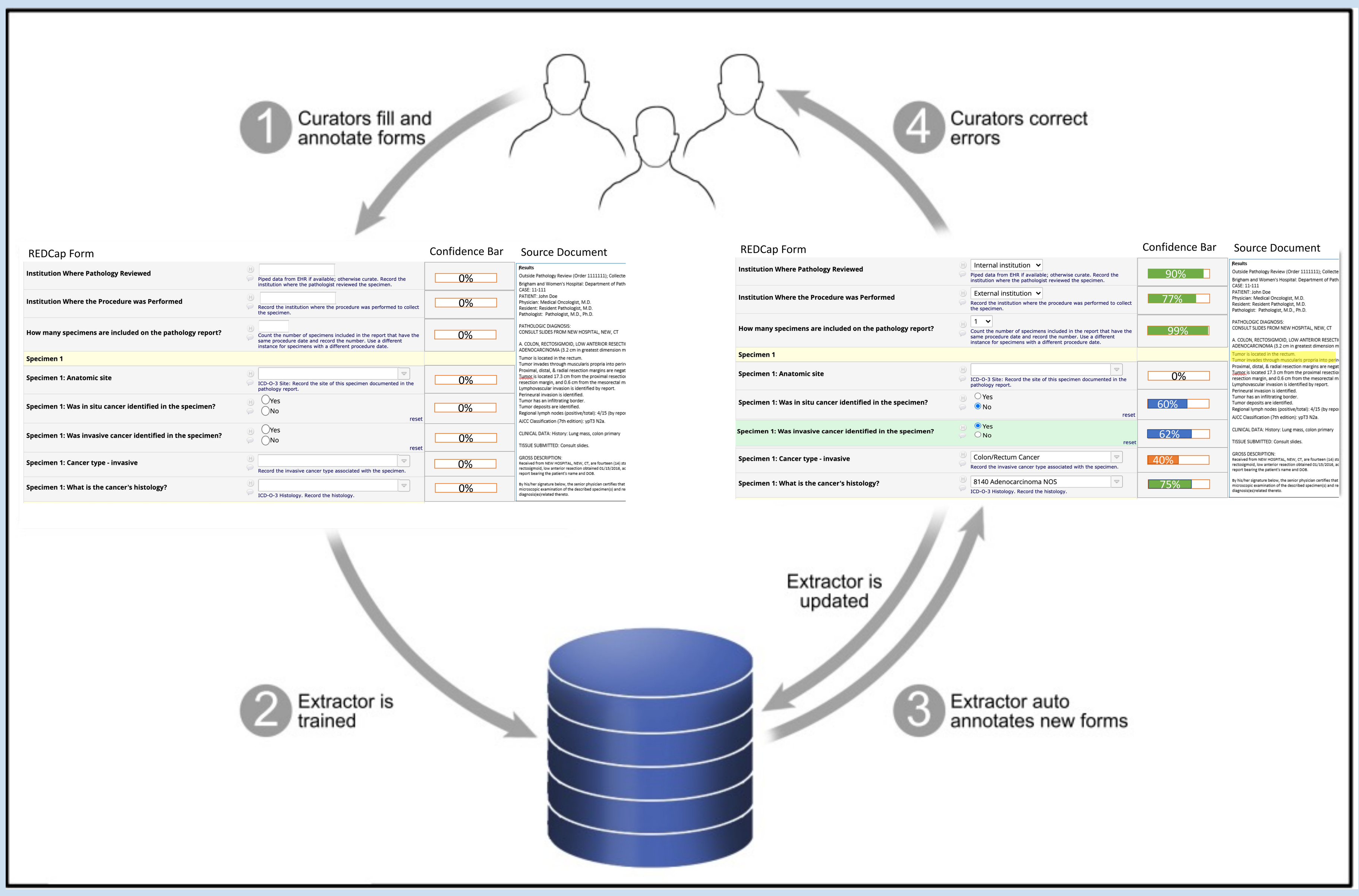
Automatic Data Curation from Unstructured Text

Protiva Rahman, PhD, Daniel Fabbri, PhD
Biomedical Informatics, Vanderbilt University Medical Center

Introduction

- Informatics research analysis requires structured data
- However, free-text documents contain valuable information
  - Electronic Health Records (EHR)
  - Biomedical Literature
- Significant time and effort spent in manual data curation
- Curators fill structured forms (e.g. REDCap) from free-text
- Existing tools do not fit into curators' workflow
  - Require additional annotation
  - Tailored for a single task (e.g. extracting gene)
- Need semi-automated tool that accelerates data curation
  - Extract and autofill form fields from free-text
  - Improve performance based on curator feedback
- We present preliminary results of our extraction model

Workflow



- Our model **extracts form fields** from EHR notes with *86% accuracy*
- **Augmenting training data** with synonym replacement *improves F1 score*
- **Focusing** on relevant region *decreases model training time* without impacting accuracy
- Incorporating **our extraction model** into a curation tool, e.g.,REDCap, will **significantly accelerate data curation** and informatics research

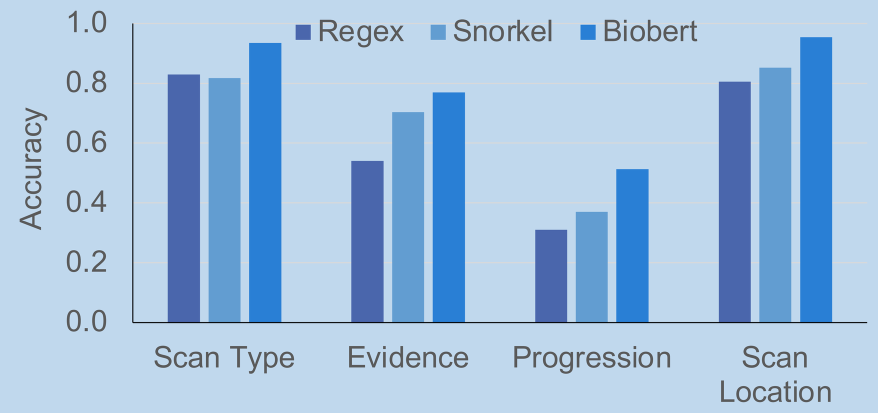
Methods

- Extracting each form field is a classification problem
  - Input: Text and form field
  - Output: Classes correspond to form field values
  - Multi-value fields (cancer sites, genes) are binarized
- Regular expression (regex): Baseline for extractions
  - Each regex rule has a different accuracy
- Snorkel[1] : Model to estimate accuracy of each regex rule
  - Rules weighted according to estimated accuracy
  - Augment training data using synonyms
- BERT[2] : State-of-the-art NLP classification model
  - Compare different BERT models
  - Performance on augmented dataset
- Focused extraction: BERT takes max input of 512 words
  - Longer text is split into multiple inputs
  - Performance of zooming in on specific region
  - Input sizes of 100, 250, 512

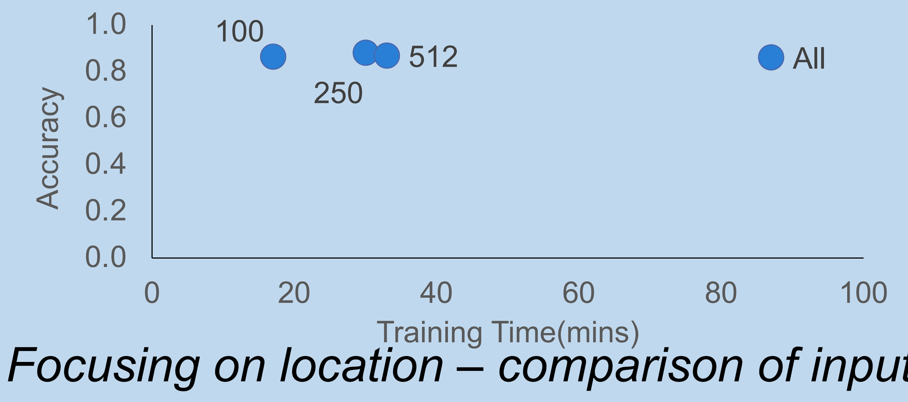
Results

Dataset	BERT (Raw)	BERT (Finetuned)	ClinicalBERT
EHR Notes	21.8	85.3	**86.8**
Biomedical Literature	30.8	58.7	**59.2**

BERT Accuracy Comparison: Finetuning BERT provides significant improvement over Google's pretrained model. Slight improvement upon using EHR trained BERT



Model Comparison: BERT outperforms regular expressions and Snorkel



Focusing on location – comparison of input sizes (labels): Minimal loss in accuracy for 75% decrease in compute time

Form Field		Original	Augmented
Scan Type		0.64	0.64
Cancer Evidence		0.51	0.47
Progression		*0.44*	*0.64*
Scan Location	Brain	0.98	0.98
	Spine	*0.66*	*0.90*
	Neck	0.67	0.67
	Chest	*0.80*	*0.86*
	Abdomen	0.86	0.87
	Pelvis	0.86	0.90
	Extremity	0.67	0.70
	Body	0.92	0.93
Average		**0.74**	**0.79**

Augmented Dataset: Significant increase in F1 score by increasing training set size with synonym replacement

References

1. Ratner A, Bach SH, Ehrenberg H, Fries J, Wu S, Ré C. Snorkel: Rapid training data creation with weak supervision. In Proceedings of the VLDB Endowment. International Conference on Very Large Data Bases 2017 Nov (Vol. 11, No. 3, p. 269).
2. Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. InNAACL-HLT (1) 2019 Jan 1.