

# Spark NLP for Healthcare Data Scientists

April 14-15, 2023

**Veysel Kocaman**  
Head of Data Science  
[veysel@johnsnowlabs.com](mailto:veysel@johnsnowlabs.com)



# Agenda

Day	Dur.	Topic	Notebooks	Instructor
Day-1	50 min	- Intro to John Snow Labs and Spark NLP - Healthcare NLP in Spark NLP	-	Veysel
	50 min	- Clinical Named Entity Recognition	01.0, 09.0, 01.4, 10.0	Veysel
	50 min	- Clinical Assertion Status Model	02.0	Hasham
	50 min	- Clinical Relation Extraction Model - Clinical Relation Extraction with Knowledge Graph (Neo4j) - Zero Shot Relation Extraction	03.0, 03.1, 03.2, 03.3	Hasham
Day-2	50 min	- Contextual Parser (rule based NER) - Zero Shot NER - NER Profiling - Chunk Merger	09.0, 01.4, 07.1, 10.0,	Veysel
	50 min	- Clinical Pretrained Pipelines - Various topics (text classification, date normalization, ADE)	07.0, 22.0, 08.0, 08.4	Veysel
	50 min	- Clinical Entity Resolution - ChunkMapper	05.0, 05.1, 05.2, 05.3, 06.0	Muhammed
	50 min	- De-Identification and Obfuscation of PHI	04.0, 04.1	Muhammed

# Colab Notebooks

## RUNNING CODE:

<https://github.com/JohnSnowLabs/spark-nlp-workshop/tree/master/healthcare-nlp>

 master / spark-nlp-workshop / healthcare-nlp /

### ► Clinical Named Entity Recognition

01.0.Clinical\_Named\_Entity\_Recognition\_Model.ipynb

01.1.prepare\_CoNLL\_from\_annotations\_for\_NER.ipynb

01.2.Resume\_MedicalNer\_Model\_Training.ipynb

01.3.BertForTokenClassification\_NER\_SparkNLP\_with\_Transformers.ipynb

01.4.ZeroShot\_Clinical\_NER.ipynb

### ► Clinical Assertion

02.0.Clinical\_Assertion\_Model.ipynb

02.1.Scope\_window\_tuning\_assertion\_status\_detection.ipynb

### ► Clinical Relation Extraction

03.0.Clinical\_Relation\_Extraction.ipynb

03.1.Clinical\_Relation\_Extraction\_BodyParts\_Models.ipynb

03.2.Clinical\_RE\_Knowledge\_Graph\_with\_Neo4j.ipynb

03.3.ZeroShot\_Clinical\_Relation\_Extraction.ipynb

03.4.Resume\_RelationExtractionApproach\_Training.ipynb

### ► Clinical De-identification

04.0.Clinical\_Deidentification.ipynb

04.1.Clinical\_Multi\_Language\_Deidentification.ipynb

04.2.Clinical\_Deidentification\_SparkNLP\_vs\_SpaCy\_vs\_Scrubadub\_vs\_Presidio\_Comparison.ipynb

04.3.Clinical\_Deidentification\_SparkNLP\_vs\_Cloud\_Providers\_Comparison.ipynb

04.4.Clinical\_Deidentification\_Improvement.ipynb

04.5.Clinical\_Deidentification.Utility\_Module.ipynb

spark-nlp==4.3.2  
spark-nlp-jsl==4.3.2

## BOOKMARK:

<https://nlp.johnsnowlabs.com/models>

<https://nlp.johnsnowlabs.com/docs/en/quickstart>

<spark-nlp.slack.com>

Go to file

Add file

...



Open in Colab

### ► Clinical Entity Resolution

05.0.Clinical\_Entity\_Resolvers.ipynb

05.1.Clinical\_Entity\_Resolver\_Model\_Training.ipynb

05.2.Finetuning\_Clinical\_Entity\_Resolver\_Model.ipynb

05.3.Calculate\_Medicare\_Risk\_Adjustment\_Score.ipynb

05.4.Sentence\_Entity\_Resolvers\_with\_EntityChunkEmbeddings.ipynb

05.5.Improved\_Entity\_Resolvers\_in\_SparkNLP\_with\_sBert.ipynb

05.6.Improved\_Entity\_Resolution\_with\_SentenceChunkEmbeddings.ipynb

### ► Entity Code Mapping

06.0.Chunk\_Mapping.ipynb

06.1.Code\_Mapping\_Pipelines.ipynb

### ► Clinical Pipelines

07.0.Pretrained\_Clinical\_Pipelines.ipynb

07.1.Pretrained\_NER\_Profiling\_Pipelines.ipynb

### ► Classifiers

08.0.Clinical\_Text\_Classification\_with\_SparkNLP.ipynb

08.1.Text\_Classification\_with\_DocumentMLClassifier.ipynb

08.2.Generic\_Classifier.ipynb

08.3.MedicalBertForSequenceClassification\_in\_SparkNLP.ipynb

08.4.Gender\_Classifier.ipynb

### ► Utility Notebooks

09.0.Contextual\_Parser\_Rule\_Based\_NER.ipynb

10.0.Clinical\_NER\_Chunk\_Merger.ipynb

11.0.SentenceDetectorDL\_Healthcare.ipynb

12.0.Clinical\_Context\_Spell\_Checker.ipynb

### ► Normalizers

13.0.Date\_Normalizer.ipynb

14.0.Drug\_Normalizer.ipynb

### ► Auxillary Notebooks

15.0.EntityRuler\_with\_Clinical\_NER\_Models.ipynb

16.0.Coreference\_Resolution\_with\_Clinical\_NER\_Models.ipynb

17.0.Graph\_builder\_for\_DL\_models.ipynb

18.0.Chunk\_Sentence\_Splitter.ipynb

19.0.Chunk\_Key\_Phrase\_Extraction.ipynb

20.0.Named\_Entity\_Disambiguation.ipynb

### ► Topic Based Notebooks

21.0.Oncology\_Model.ipynb

22.0.Adverse\_Drug\_Event\_ADE\_NER\_and\_Classifier.ipynb

23.0.Medical\_Question\_Answering

## Spark NLP in Action

Spark NLP - English → Recognize Entities

Spark NLP, English

- Infer Meaning & Intent
- Classify Sentiments
- Recognize Entities
- Detect Sentiment & Emotion
- Analyze Spelling & Grammar

Spark NLP, World Languages

- Identify & Translate Languages
- European Languages
- East Asian Languages
- Language of India
- Middle Eastern Languages
- Language of Africa

Spark NLP for Healthcare

- Recognize Clinical Entities
- Recognize Biomedical Entities
- De-identification
- Resolve Entities to Codes
- Recognize Social Determinants
- Extract Entities
- Soft & Clean Medical Text
- Analyze non-English Medical Text

Spark OCR

- Extract Text from Documents
- Enhance Low-Quality Images
- Extract Tables & Structured Data
- Analyze non-English Medical Text

Home Docs Learn Models Demo 🚀⚙️

## John Snow Labs NLP Documentation



## NLP Models Hub

A place for sharing and discovering Spark NLP models and pipelines

Search models and pipelines

Show All models & pipelines in All Languages for All Spark NLP versions

4,390 Models & Pipelines Results:

Model Type	Description	Date	Task	Language	Edition
SUPPORTED	Sentence Entity Resolver for UMLS CUI Codes (Clinical Drugs)	10/2021	Entity Resolution	English	Spark NLP for Healthcare 3.2.3
SUPPORTED	Sentence Entity Resolver for UMLS CUI Codes (Disease or Syndrome)	10/2021	Entity Resolution	English	Spark NLP for Healthcare 3.2.3
SUPPORTED	Sentence Entity Resolver for RxNorm (product_base_cased_mit embeddings)	10/2021	Entity Resolution	English	Spark NLP for Healthcare 3.2.3
SUPPORTED	Longformer Token Classification Base - NEB CoNLL	10/2021	Named Entity Recognition	English	Spark NLP for Healthcare 3.2.3
SUPPORTED	Longformer Token Classification Base - NEB CoNLL	10/2021	Named Entity Recognition	English	Spark NLP for Healthcare 3.2.3
SUPPORTED	Sentence Entity Resolver for RxNorm (NDC)	10/2021	Entity Resolution	English	Spark NLP for Healthcare 3.2.3

Code Issues Pull requests Discussions Actions Projects Wiki Security

master · 75 branches · 7 tags · Go to file · Add file · Code ·

galiph Merge pull request #400 from JohnSnowLabs/galiph · 2 hours ago · 1,601 commits

- data Add files via upload
- dbnabrics Update benchmark.md
- java add java examples
- jupyter Update docker-compose.yaml
- mlu mlu notebooks updated
- platforms added more info for troubleshooting
- scala update scala codes
- tutorials Notebooks updated with v3.3.0
- zeppelin clean notebook
- .gitattributes Ignore html from linguist-vendorized
- .gitignore removed outdated notebooks
- ISSUE\_TEMPLATE.md Create ISSUE\_TEMPLATE.md
- LICENSE Initial commit
- README.md Update README.md
- colab\_setup.sh Update colab\_setup.sh
- jl\_colab\_setup.sh Update jl\_colab\_setup.sh
- jl\_colab\_setup\_with\_OCR.sh Update jl\_colab\_setup\_with\_OCR.sh
- jl\_sagemaker\_setup.sh Corrected paths for spark binaries
- jl\_sagemaker\_setup\_3.0.tsh added script
- jl\_sagemaker\_setup\_with\_OCR.sh Update jl\_sagemaker\_setup\_with\_OCR.sh

com.johnsnowlabs.nlp.annotator

**C** **MedicalNerApproach** Companion object **MedicalNerApproach**

```
class MedicalNerApproach extends AnnotatorApproach[MedicalNerModel] with NerApproach[MedicalNerAnnotator] with Logging with ParamAndFeatureWriterTrait with DefaultLicense
```

Linear Superatypes

Filter all members

**anno**

```
val inspectionType: Array[String]
Input annotator types: DOCUMENT, TOKEN, WORD_EMBEDDINGS

val outputAnnotatorType: String
Input annotator types: NAMED_ENTITY
```

**getParam**

```
def getBatchSize: Int
Batch size

def getConfigProtoBytes: Option[Array[Byte]]
Config proto from tensorflow, serialized into byte array.

def getDropout: Float
Dropout coefficient

def getEnableNerOptimizer: Boolean
Memory Optimizer

def getEnableOutputLogs: Boolean
Whether to output to annotators log folder

def getIncludeAllConfidenceScores: Boolean
```

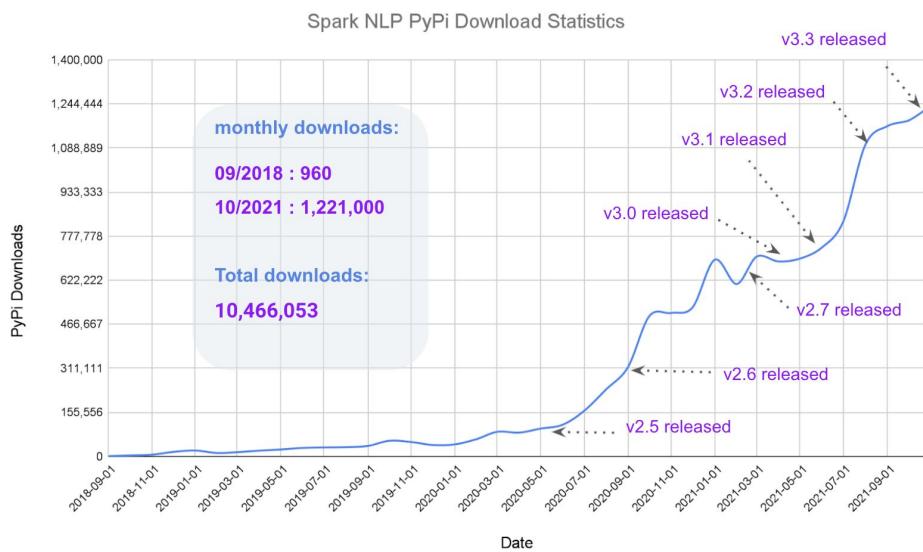
# Part - I

- ❖ Overview and key concepts in Spark NLP
- ❖ NLP basics & review
- ❖ Common medical NLP use cases
- ❖ Clinical named entity recognition

# Introducing Spark NLP

Daily ~ 100K  
Monthly ~ 2.8M

PyPI link	<a href="https://pypi.org/project/spark-nlp">https://pypi.org/project/spark-nlp</a>
Total downloads	46,372,442
Total downloads - 30 days	2,793,444
Total downloads - 7 days	667,029



- Spark NLP is an open-source natural language processing library, built on top of Apache Spark and Spark ML. (initial release: Oct 2017)
  - A single unified solution for all your NLP needs
  - Take advantage of transfer learning and implementing the latest and greatest SOTA algorithms and models in NLP research
  - The most widely used NLP library in industry (3 yrs in a row) - downloaded 46+ million times !
  - Delivering a mission-critical, enterprise grade NLP library (used by multiple Fortune 500)
  - Full-time development team (a new release every other week)

# TRUSTED BY



Imperial College  
London



STANFORD  
UNIVERSITY

# Spark NLP for Healthcare

Spark NLP for Healthcare provides

- accurate,
- scalable,
- private,
- tunable,
- modular

software library that helps healthcare & pharma organizations build longitudinal patient records and knowledge graphs on real-world EHR data.

Clinical Entity Recognition	Clinical Entity Linking	Assertion Status	Relation Extraction						
<p>40 units <b>DOSAGE</b> of insulin glargine <b>DRUG</b> at night <b>FREQUENCY</b></p>	<p>Suspect diabetes SNOMED-CT: <b>473127005</b> Lisinopril 10 MG RxNorm: <b>316151</b> Hyponatremia ICD-10: <b>E87.1</b></p>	<p>Fever and sore throat → PRESENT No stomach pain → ABSENT Father with Alzheimer → FAMILY</p>							
Algorithms		Content							
<p><b>Extract Knowledge</b></p> <ul style="list-style-type: none"> <li>• Entity Linker</li> <li>• Entity Disambiguator</li> <li>• Document Classifier</li> <li>• Contextual Parser</li> </ul>		<p><b>De-identify text</b></p> <ul style="list-style-type: none"> <li>• Structured Data</li> <li>• Unstructured Text</li> <li>• Obfuscator</li> <li>• Generalizer</li> </ul>							
<p><b>Split Text</b></p> <ul style="list-style-type: none"> <li>• Sentence Detector</li> <li>• Deep Sentence Detector</li> <li>• Tokenizer</li> <li>• nGram Generator</li> </ul>		<p><b>Clean Medical Text</b></p> <ul style="list-style-type: none"> <li>• Spell Checking</li> <li>• Spell Correction</li> <li>• Normalizer</li> <li>• Stopword Cleaner</li> </ul>							
<p><b>Clinical Grammar</b></p> <ul style="list-style-type: none"> <li>• Stemmer</li> <li>• Lemmatizer</li> <li>• Part of Speech Tagger</li> <li>• Dependency Parser</li> </ul>		<p><b>Find in Text</b></p> <ul style="list-style-type: none"> <li>• Text Matcher</li> <li>• Regex Matcher</li> <li>• Date Matcher</li> <li>• Chunker</li> </ul>							
Trainable & Tunable	Scalable to a Cluster	Fast Inference	Hardware Optimized						
			 						
Community									
<a href="#">Get Started</a>		<a href="#">View Documentation</a>							
<p><b>800+ Pretrained Models</b></p> <table border="1"> <tr> <td><b>Clinical:</b> Signs, Symptoms, Treatments, Procedures, Tests, Labs, Sections</td> <td><b>Anatomy:</b> Organ, Subdivision, Cell, Structure Organism, Tissue, Gene, Chemical</td> </tr> <tr> <td><b>Drugs:</b> Name, Dosage, Strength, Route, Duration, Frequency Poisons, Adverse Effects</td> <td><b>Demographics:</b> Age, Gender, Height, Weight, Race, Ethnicity, Marital Status, Vital Signs</td> </tr> <tr> <td><b>Risk Factors:</b> Smoking, Obesity, Diabetes, Hypertension, Substance Abuse</td> <td><b>Sensitive Data:</b> Patient Name, Address, Phone, Email, Dates, Providers, Identifiers</td> </tr> </table>				<b>Clinical:</b> Signs, Symptoms, Treatments, Procedures, Tests, Labs, Sections	<b>Anatomy:</b> Organ, Subdivision, Cell, Structure Organism, Tissue, Gene, Chemical	<b>Drugs:</b> Name, Dosage, Strength, Route, Duration, Frequency Poisons, Adverse Effects	<b>Demographics:</b> Age, Gender, Height, Weight, Race, Ethnicity, Marital Status, Vital Signs	<b>Risk Factors:</b> Smoking, Obesity, Diabetes, Hypertension, Substance Abuse	<b>Sensitive Data:</b> Patient Name, Address, Phone, Email, Dates, Providers, Identifiers
<b>Clinical:</b> Signs, Symptoms, Treatments, Procedures, Tests, Labs, Sections	<b>Anatomy:</b> Organ, Subdivision, Cell, Structure Organism, Tissue, Gene, Chemical								
<b>Drugs:</b> Name, Dosage, Strength, Route, Duration, Frequency Poisons, Adverse Effects	<b>Demographics:</b> Age, Gender, Height, Weight, Race, Ethnicity, Marital Status, Vital Signs								
<b>Risk Factors:</b> Smoking, Obesity, Diabetes, Hypertension, Substance Abuse	<b>Sensitive Data:</b> Patient Name, Address, Phone, Email, Dates, Providers, Identifiers								

# Academic Activities & Benchmarks



## Preparing for the Next Pandemic: Transfer Learning from Existing Diseases via Hierarchical Multi-Modal BERT Models to Predict COVID-19 Outcomes

Khushbu Agarwal<sup>1</sup>, Sutanay Choudhury<sup>1\*</sup>, Sindhu Tipirneni<sup>2</sup>, Pritam Mukherjee<sup>3</sup>, Colby Ham<sup>1</sup>, Suzanne Tamang<sup>1</sup>, Matthew Baker<sup>4</sup>, Siyi Tang<sup>5</sup>, Veysel Kocaman<sup>7</sup>, Olivier Gervais<sup>1,2,4</sup>, Robert Rallo<sup>1</sup>, and Chandan K Reddy<sup>1</sup>

<sup>1</sup>Pacific Northwest National Laboratory, Richland, 99354, USA

<sup>2</sup>Department of Computer Science, Virginia Tech, Arlington, 22203, USA

<sup>3</sup>Stanford Center for Biomedical Informatics Research, Department of Medicine, School of Medicine, Stanford University, Stanford, 94305, USA

<sup>4</sup>Department of Biomedical Data Science, Stanford University, Stanford, 94305, USA

<sup>5</sup>Department of Electrical Engineering, Stanford University, Stanford, 94305, USA

<sup>6</sup>Division of Immunology and Rheumatology, Department of Medicine, Stanford University, Stanford, 94305, USA

<sup>7</sup>John Snow Labs, Delaware City, 19968, USA

stanford.cs@stanford.edu



INFORMATICS PROFESSIONALS. LEADING THE WAY.

American Medical  
Informatics  
Association

## Tracking the Evolution of COVID-19 via Temporal Comorbidity Analysis from Multi-Modal Data

Sutanay Choudhury<sup>1</sup>, Khushbu Agarwal<sup>1</sup>, Colby Ham<sup>1</sup>, Pritam Mukherjee<sup>2</sup>, Siyi Tang<sup>3</sup>, Sindhu Tipirneni<sup>3</sup>, Chandan Reddy<sup>4</sup>, Suzanne Tamang<sup>2</sup>, Robert Rallo<sup>1</sup>, Veysel Kocaman<sup>7</sup>,  
<sup>1</sup>Pacific Northwest National Laboratory; <sup>2</sup>Stanford University; <sup>3</sup>Virginia Tech;

John Snow Labs

### Introduction

We aim to characterize the evolution in the effectiveness of treatment for different patient groups over the course of the COVID-19 pandemic. In contrast to most existing studies<sup>1</sup>, we study the evolution of patient trajectories based on unique sets of frequent comorbid conditions discovered from the data. Further, we study the association between frequent co-morbid conditions to the length of stay (LOS) as a measure of treatment efficacy, for poor COVID-19 related outcomes.

### Journal of Biomedical Semantics

#### SOFTWARE

## Accurate Clinical and Biomedical Named Entity Recognition at Scale

Veysel Kocaman\* and David Talby

\*Correspondence:  
veysel@johnsnowlabs.com  
John Snow Labs, Lewes, DE, USA  
Full list of author information is available at the end of the article

### Scientific Document Understanding (SDU) at AAAI

## Deeper Clinical Document Understanding Using Relation Extraction

Hasham Ul Haq, Veysel Kocaman, David Talby

John Snow Labs Inc.  
16192 Coastal Highway  
Lewes, DE, USA 19958  
{hasham, veysel, david}@johnsnowlabs.com

#### Abstract

The surging amount of biomedical literature & digital clinical records presents a growing need for text mining techniques that can not only identify but also semantically enrich

publications and literature are growing rapidly, there still lacks structured knowledge that can be easily processed by computer programs. Relation Extraction becomes even more pertinent in biomedical research as it can provide the criti-



## New State-of-the-art (SOTA) Benchmarks



- ✓ 6 academic publications & events and 1 patent application, 20+ medium blogposts
- ✓ new SOTA benchmarks on Clinical NER challenges (i2b2 2010 Clinical, i2b2 2014 Deid, n2c2 2018 Medication)
- ✓ new SOTA benchmarks on Adverse Drug Reaction NER datasets (ADE, CADEC, SMM4H )
- ✓ new SOTA benchmarks on Adverse Drug Reaction classification datasets (ADR, CADEC)
- ✓ new SOTA benchmarks on Clinical Relation Extraction datasets (i2b2, temporal, ADE, Posology, PGR – 5 out of 7)



Health  
Intelligence  
(W3PHIAI-22)  
at AAAI

### Mining Adverse Drug Reactions from Unstructured Mediums at Scale

Hasham Ul Haq, Veysel Kocaman, David Talby

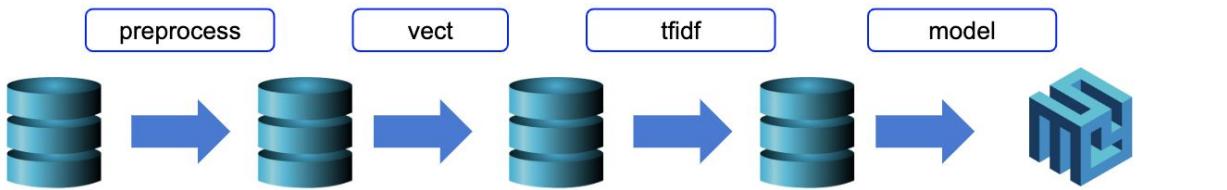
John Snow Labs Inc.  
16192 Coastal Highway  
Lewes, DE, USA 19958  
{hasham, veysel, david}@johnsnowlabs.com

ADR's has been estimated to cost \$156 billion each year in the United States alone (van Der Hoof et al. 2006).

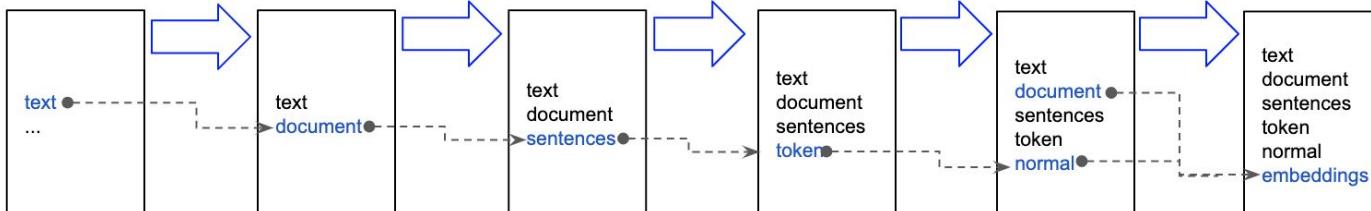
Finding all ADR's of a drug before it is marketed is not practical for several reasons. First, the number of human subjects going through clinical trials is often too small to detect rare ADR's. Second, many clinical trials are short-lasting while some ADR's take time to manifest. Third,

# Introducing Spark NLP

## Pipeline of annotators



DocumentAssembler() SentenceDetector() Tokenizer() Normalizer() WordEmbeddings()



DataFrame

```
from pyspark.ml import Pipeline
document_assembler = DocumentAssembler()\
    .setInputCol("text")\
    .setOutputCol("document")
sentenceDetector = SentenceDetector()\
    .setInputCols(["document"])\
    .setOutputCol("sentences")
tokenizer = Tokenizer() \
    .setInputCols(["sentences"]) \
    .setOutputCol("token")
normalizer = Normalizer()\
    .setInputCols(["token"])\
    .setOutputCol("normal")
word_embeddings=WordEmbeddingsModel.pretrained()\
    .setInputCols(["document","normal"])\
    .setOutputCol("embeddings")
nlpPipeline = Pipeline(stages=[document_assembler,
    sentenceDetector,
    tokenizer,
    normalizer,
    word_embeddings,
])
nlpPipeline.fit(df).transform(df)
```

# Introducing Spark NLP



Faster inference

```
from sparknlp.base import LightPipeline  
LightPipeline(someTrainedPipeline).annotate(someStringOrArray)
```

Spark is like a [locomotive](#) racing a [bicycle](#). The [bike](#) will win if the load is light, it is quicker to accelerate and more agile, but with a heavy load the [locomotive](#) might take a while to get up to speed, but [it's](#) going to be faster in the end.

**LightPipelines** are Spark ML pipelines converted into a single machine but multithreaded task, becoming more than 10x times faster for smaller amounts of data (small is relative, but 50k sentences is roughly a good maximum).

# Healthcare NLP

Clean & structured data



Raw & unstructured data



Healthcare data



- Less than **50% of the structured data** and less than **1% of the unstructured data** is being leveraged for decision making in companies (HBR). This is even worse in healthcare.
- NLP is ultra domain specific, so train your own models.

# Why is language understanding hard?

## Human Language is:

- Nuanced
- Fuzzy
- Contextual
- Medium specific
- Domain specific

## Healthcare specific needs:

### 1. Core Annotators

Part of speech, spell checking, ...

### 2. Vocabulary

Ontologies, relationships, word embeddings, ...

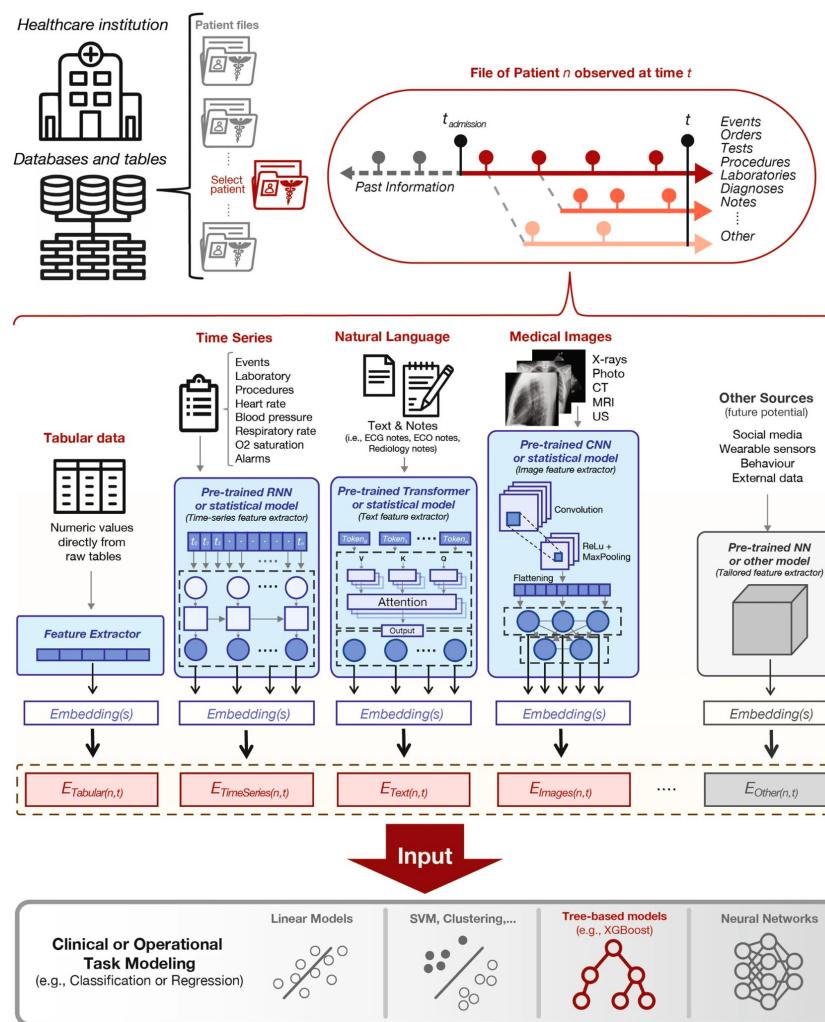
### 3. ML & DL Models

Named entity recognition, entity resolution, ...

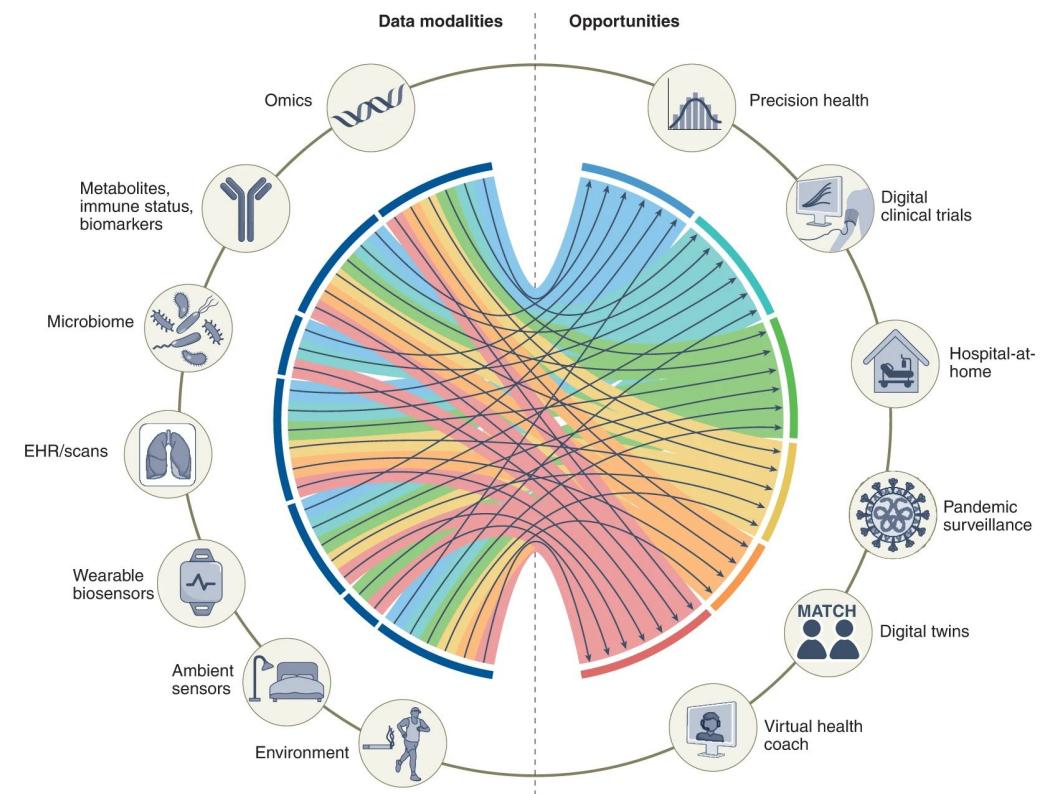
ED Triage Notes
states started last night, upper abd, took alka seltzer approx 0500, no relief. nausea no vomiting
Since yesterday 10/10 "constant Tylenol 1 hr ago. +nausea. diaphoretic. Mid abd radiates to back
Generalized abd radiating to lower x 3 days accompanied by dark stools. Now with bloody stool this am. Denies dizzy, sob, fatigue. Visiting from Japan on business."



Features	
Type of Pain	Symptoms
Intensity of Pain	Onset of symptoms
Body part of region	Attempted home remedy



# Multimodality in Healthcare



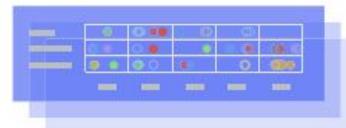
# Healthcare NLP

Unstructured EHR  
data



Models that work at hospital 1 don't work at hospital 2

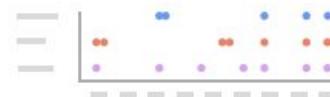
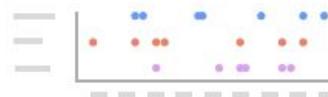
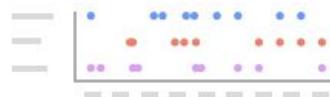
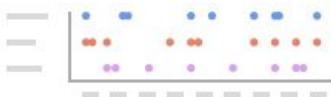
Map to common  
format (FHIR)



Fast Healthcare  
Interoperability  
Resources

Models can be used across health systems

Temporal  
sequencing



Sequence models have access to the entire patient's record

Use AI to answer  
questions

What parts of a patient's past  
history should be reviewed?

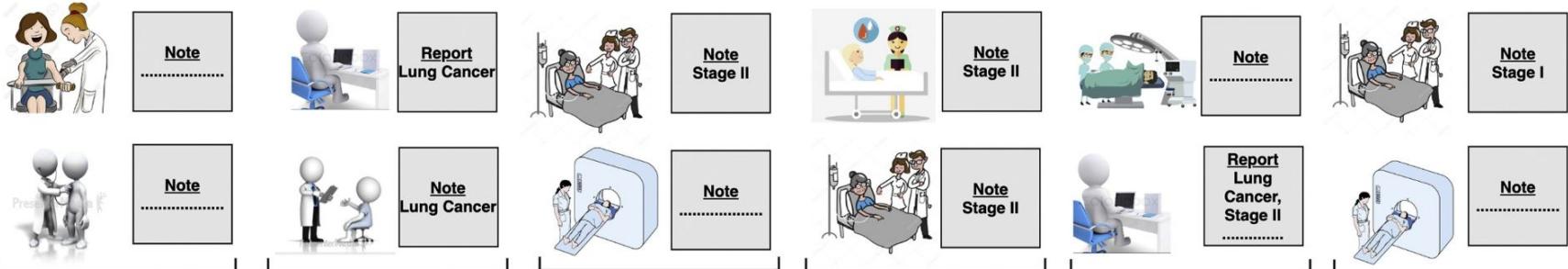
What about a patient's current  
state needs to be known?

What are opportunities  
to intervene?

What are the risks  
of future outcomes?

# Putting the clinical facts on a timeline

Natural History



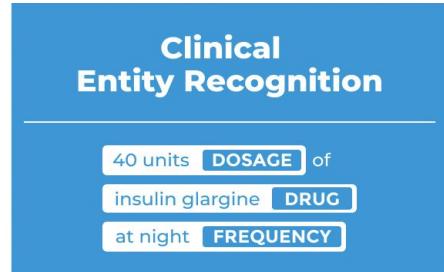
Medical Timeline

Lung Cancer  
Diagnosis

Tumor Stage II

Tumor Stage I

# NLP in Healthcare



**Clinical Entity Linking**

Suspect diabetes SNOMED-CT: 473127005

Lisinopril 10 MG RxNorm: 316151

Pyponatremia ICD-10: E87.1

**Assertion Status**

Fever and sore throat → PRESENT

No stomach pain → ABSENT

Father with Alzheimer → FAMILY

**De-Identification**

Ora **NAME**, a **25 AGE** yo  
cashier **PROFESSION** from  
Morocco **LOCATION**

**Relation Extraction**

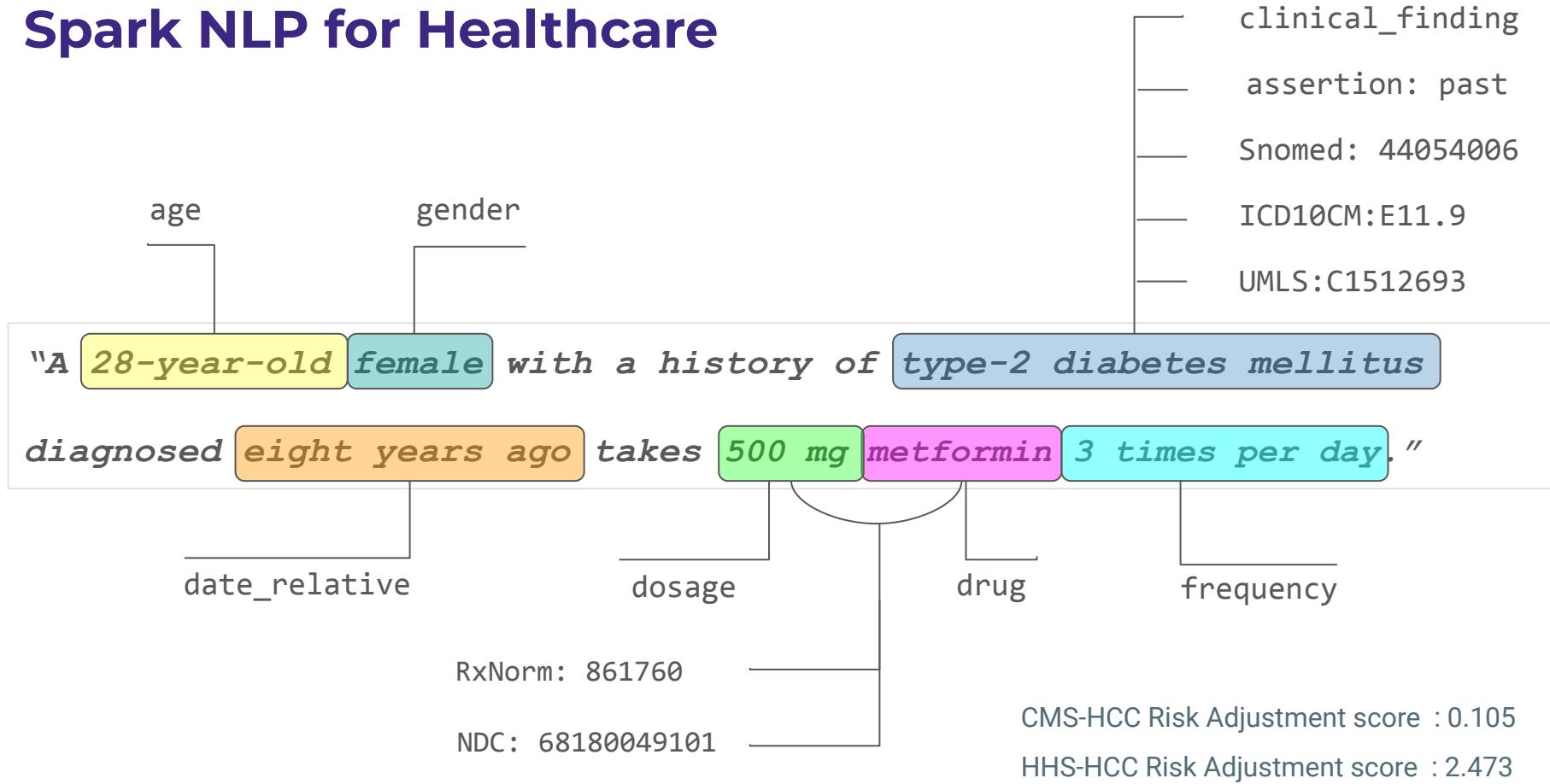
AFTER

Admitted for **nausea** due to **chemo**

Occurrence Symptom Treatment

CAUSED BY

# Spark NLP for Healthcare



# Spark NLP for Healthcare

Named Entity Recognition

ICD10 Resolver

Snomed Resolver

UMLS Resolver

Assertion Status Detection

Risk Adj. Module

RxNorm Resolver

Relationship Extraction

clinical\_finding

Snomed: 44054006

ICD10CM:E11.9

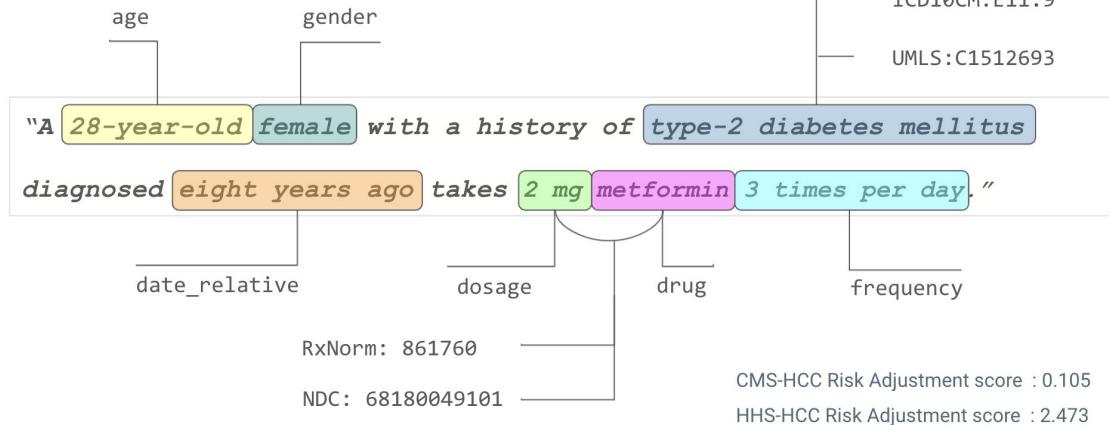
UMLS:C1512693

Sentence Splitter

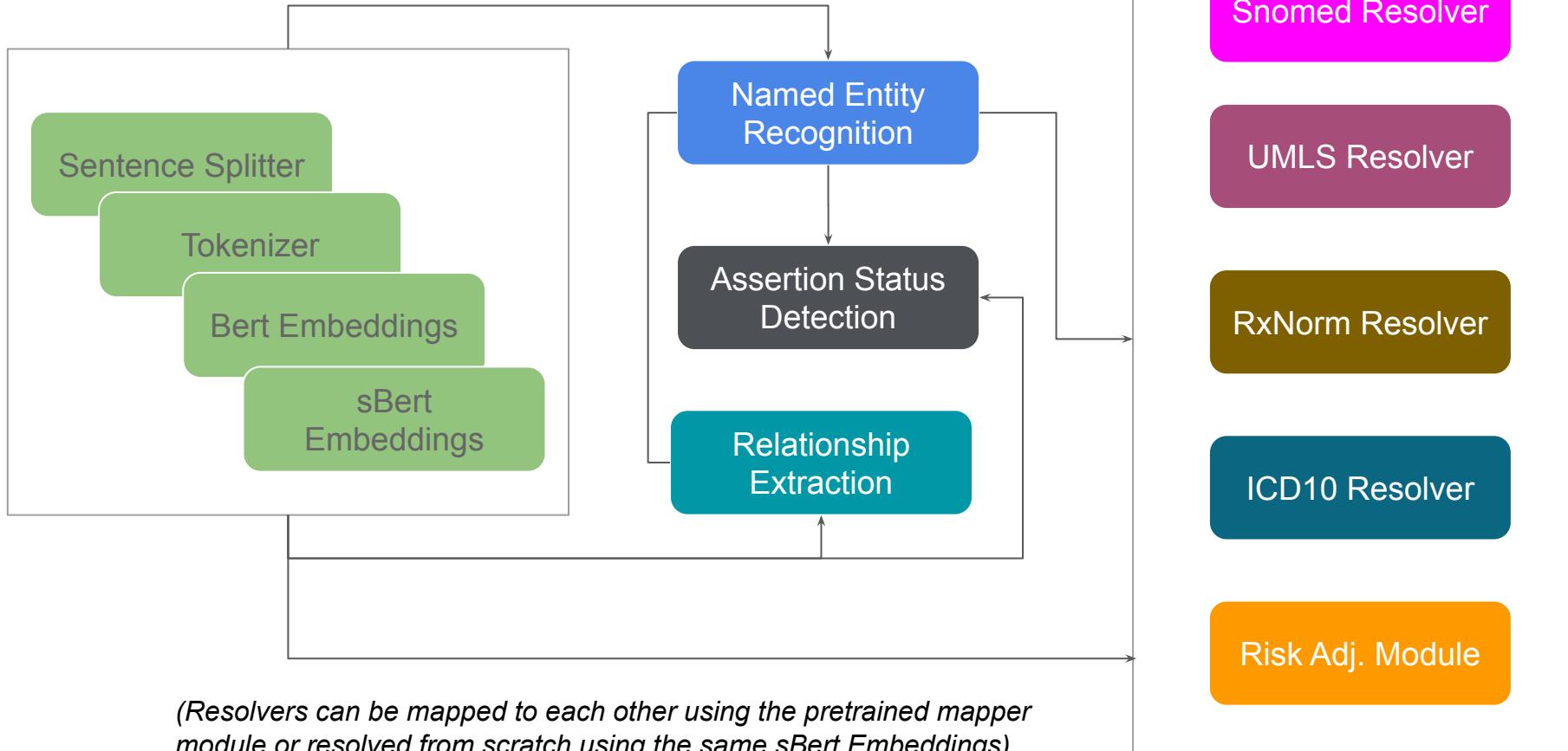
Tokenizer

Bert Embeddings

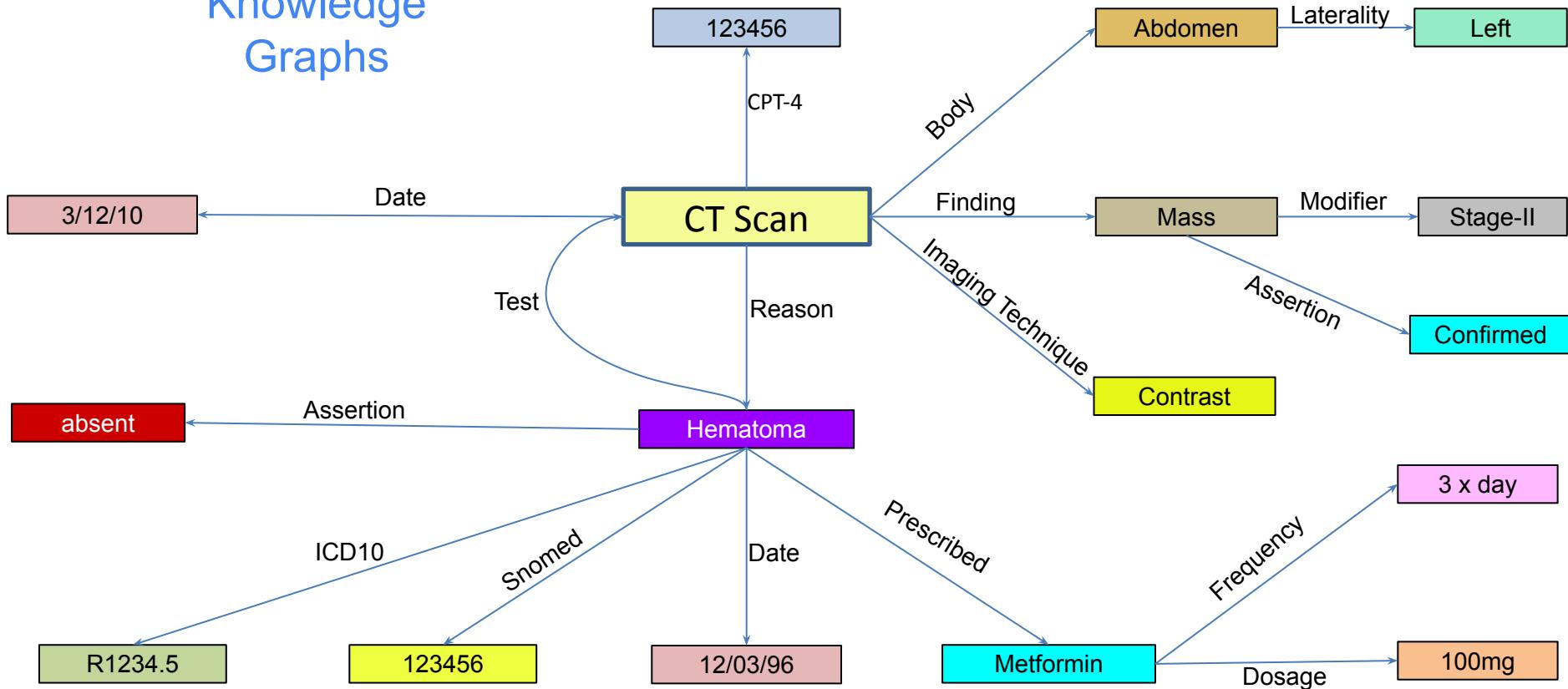
sBert Embeddings

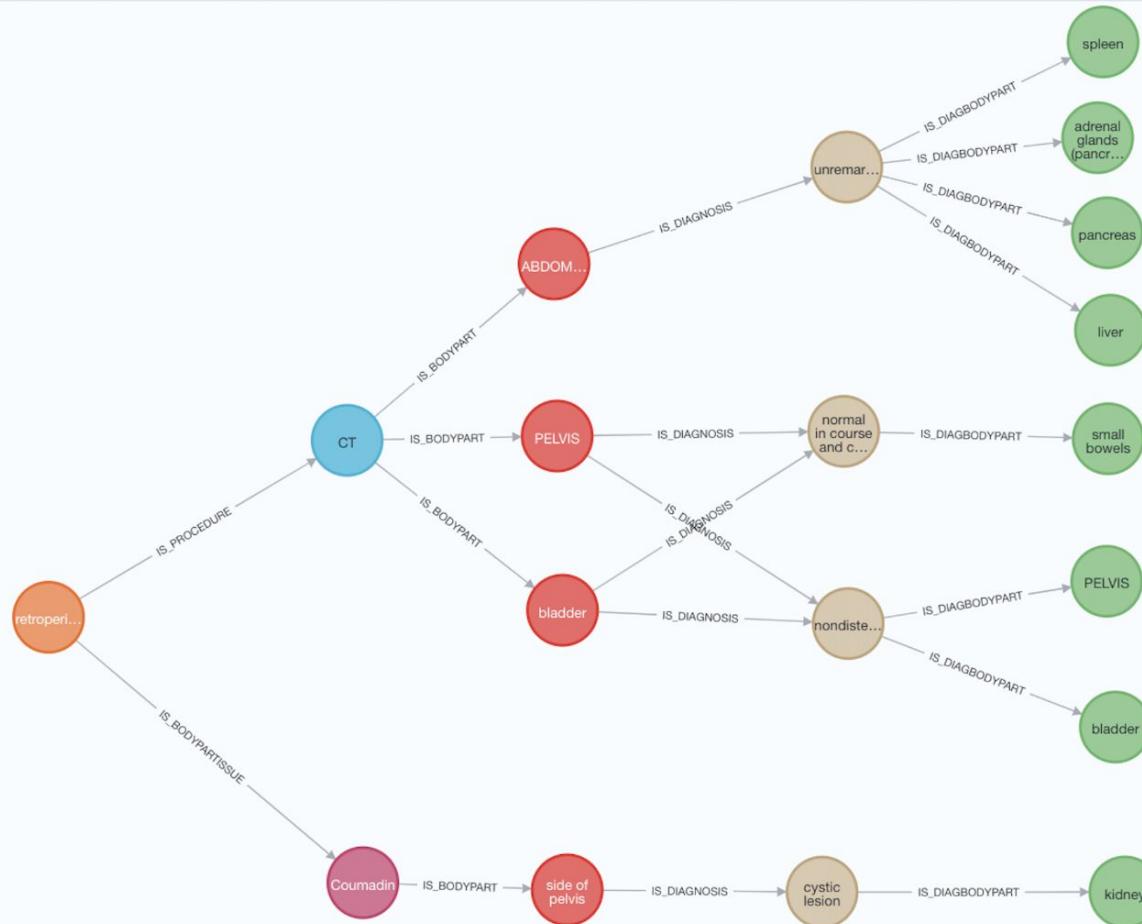


# Spark NLP for Healthcare



# Knowledge Graphs





REASON FOR EXAM: Evaluate for retroperitoneal hematoma on the right side of pelvis, the patient has been following, is currently on Coumadin.

**CT ABDOMEN:** There is no evidence for a retroperitoneal hematoma.

The liver, spleen, adrenal glands, and pancreas are unremarkable.

Within the superior pole of the left kidney, there is a 3.9 cm cystic lesion.

A 3.3 cm cystic lesion is also seen within the inferior pole of the left kidney.

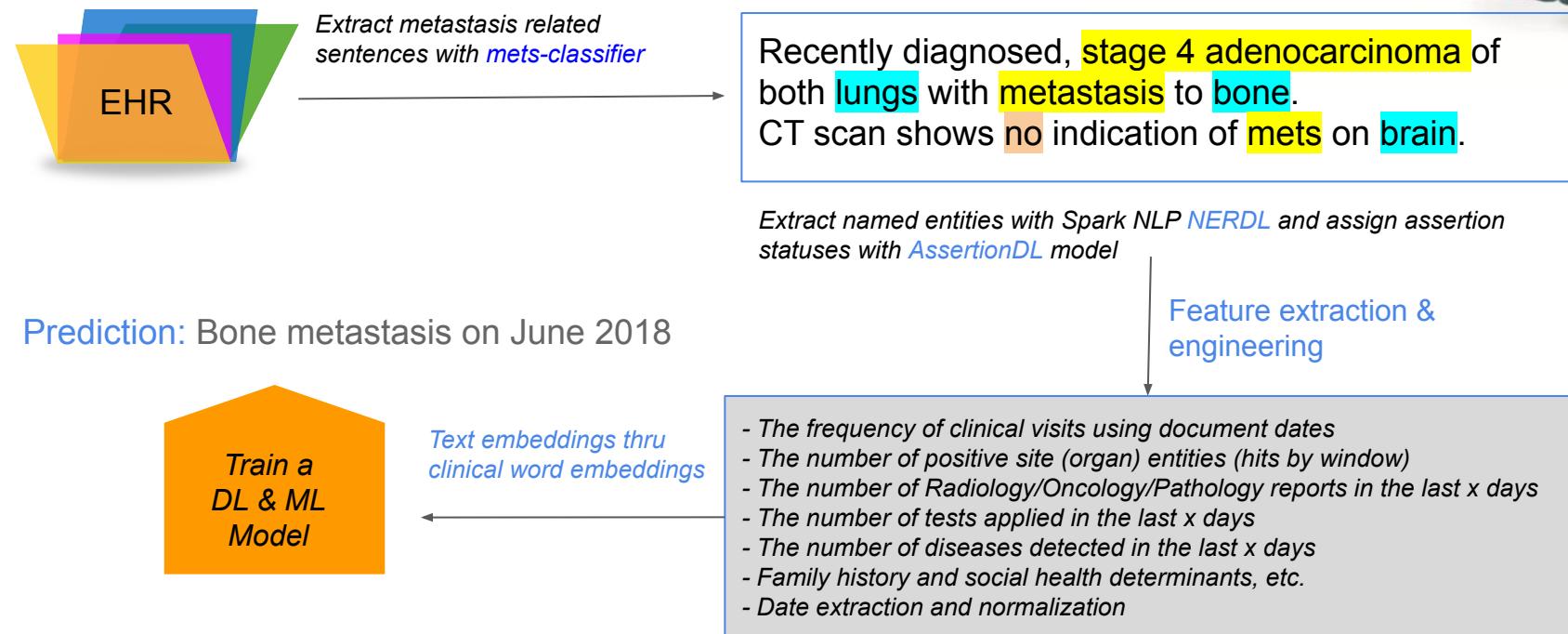
No calcifications are noted. The kidneys are small bilaterally.

**CT PELVIS:** Evaluation of the bladder is limited due to the presence of a Foley catheter, the bladder is nondistended.

The large and small bowels are normal in course and caliber. There is no obstruction.

# NLP in Healthcare

Case: Predicting if a patient would develop a metastasis on certain sites.



# Spark NLP for Healthcare - Pipeline Components

```

from pyspark.ml import Pipeline

documentAssembler = DocumentAssembler()\
    .setInputCol("text")\
    .setOutputCol("document")

sentenceDetector = SentenceDetector()\
    .setInputCols(["document"])\
    .setOutputCol("sentences")

tokenizer = Tokenizer() \
    .setInputCols(["sentences"]) \
    .setOutputCol("token")

normalizer = Normalizer()\
    .setInputCols(["token"])\
    .setOutputCol("normal")

word_embeddings=WordEmbeddingsModel.pretrained()\
    .setInputCols(["document", "normal"])\\
    .setOutputCol("embeddings")

nlpPipeline = Pipeline(stages=[\
    documentAssembler, \
    sentenceDetector, \
    tokenizer, \
    normalizer, \
    word_embeddings, \
])

nlpPipeline.fit(df).transform(df)

```



Algorithms		Content			
<b>Extract Knowledge</b>	<b>De-identify text</b>	<b>Medical Transformers</b>	<b>Linked Medical Terminologies</b>		
<ul style="list-style-type: none"> <li>Entity Linker</li> <li>Entity Disambiguator</li> <li>Document Classifier</li> <li>Contextual Parser</li> </ul>	<ul style="list-style-type: none"> <li>Structured Data</li> <li>Unstructured Text</li> <li>Obfuscator</li> <li>Generalizer</li> </ul>	<div style="display: flex; justify-content: space-around;"> <div>JSL-BERT-Clinical</div> <div>BioBERT</div> <div>ClinicalBERT</div> <div>GloVe-Med</div> <div>GloVe-ICD-O</div> <div>BlueBERT</div> </div>	<div style="display: flex; justify-content: space-around;"> <div>SNOMED-CT</div> <div>CPT</div> <div>UMLS</div> <div>ICD-10-CM</div> <div>RxNorm</div> <div>HPO</div> <div>ICD-10-PCS</div> <div>ICD-O</div> <div>LOINC</div> </div>		
<b>Split Text</b>	<b>Clean Medical Text</b>	<b>800+ Pretrained Models</b>			
<ul style="list-style-type: none"> <li>Sentence Detector</li> <li>Deep Sentence Detector</li> <li>Tokenizer</li> <li>nGram Generator</li> </ul>	<ul style="list-style-type: none"> <li>Spell Checking</li> <li>Spell Correction</li> <li>Normalizer</li> <li>Stopword Cleaner</li> </ul>	<div><b>Clinical:</b> Signs, Symptoms, Treatments, Procedures, Tests, Labs, Sections</div> <div><b>Anatomy:</b> Organ, Subdivision, Cell, Structure, Organism, Tissue, Gene, Chemical</div> <div><b>Drugs:</b> Name, Dosage, Strength, Route, Duration, Frequency, Poisons, Adverse Effects</div> <div><b>Demographics:</b> Age, Gender, Height, Weight, Race, Ethnicity, Marital Status, Vital Signs</div> <div><b>Risk Factors:</b> Smoking, Obesity, Diabetes, Hypertension, Substance Abuse</div> <div><b>Sensitive Data:</b> Patient Name, Address, Phone, Email, Dates, Providers, Identifiers</div>			
Trainable & Tunable		Scalable to a Cluster	Fast Inference	Hardware Optimized	Community

# DocumentAssembler

- ✓ Prepares data into a format that is processable by Spark NLP

"A 28-year-old female with a history  
of gestational diabetes mellitus  
diagnosed eight years prior to  
presentation and subsequent type two  
diabetes mellitus"

**text**

**DocumentAssembler**



	document	begin	end
0	A 28-year-old female with a history of gestational diabetes mellitus diagnosed eight years prior to presentation and subsequent type two diabetes mellitus	0	153

**document**

```
documentAssembler = DocumentAssembler()\\"  
.setInputCol("text")\"  
.setOutputCol("document")
```

# SentenceDetectorDLModel

- ✓ Detects sentence boundaries using a deep learning approach.
- ✓ `sentence_detector_dl_healthcare` is quite successful for clinical documents

document begin end

A 28-year-old female with a history of gestational diabetes mellitus diagnosed eight years prior to presentation and subsequent type two diabetes mellitus ( T2DM ) , one prior episode of HTG-induced pancreatitis three years prior to presentation , associated with an acute hepatitis , and obesity with a body mass index ( BMI ) of 33.5 kg/m<sup>2</sup> , presented with a one-week history of polyuria , polydipsia , poor appetite , and vomiting . Two weeks prior to presentation , she was treated with a five-day course of amoxicillin for a respiratory tract infection . She was on metformin , glipizide , and dapagliflozin for T2DM and atorvastatin and gemfibrozil for HTG . She had been on dapagliflozin for six months at the time of presentation . Physical examination on presentation was significant for dry oral mucosa ; significantly , her abdominal examination was benign with no tenderness , guarding , or rigidity . Pertinent laboratory findings on admission were : serum glucose 111 mg/dL , bicarbonate 18 mmol/L , anion gap 20 , creatinine 0.4 mg/dL , triglycerides 508 mg/dL , total cholesterol 122 mg/dL , glycated hemoglobin ( HbA1c ) 10% , and venous pH 7.27 . Serum lipase was normal at 43 U/L . Serum acetone levels could not be assessed as blood samples kept hemolyzing due to significant lipemia . The patient was initially admitted for starvation ketosis , as she reported poor oral intake for three days prior to admission . Her initial serum glucose was 188 mg/dL . After a meal , her glucose was 188 mg/dL . Her serum lipase was 42 U/L . Serum bicarbonate was 16 mmol/L . The  $\beta$ -hydroxybutyrate level was obtained and found to be elevated at 5.29 mmol/L – the original sample was centrifuged and the chylomicron layer removed prior to analysis due to interference from turbidity caused by lipemia again . The patient was treated with an insulin drip for euDKA and HTG with a reduction in the anion gap to 13 and triglycerides to 1400 mg/dL , within 24 hours . Her euDKA was thought to be precipitated by her respiratory tract infection in the setting of SGLT2 inhibitor use . The patient was seen by the endocrinology service and she was discharged on 40 units of insulin glargine at night , 12 units of insulin lispro with meals , and metformin 1000 mg two times a day . It was determined that all SGLT2 inhibitors should be discontinued indefinitely . She had close follow-up with endocrinology post discharge .

0 2515



		sentence	begin	end
0	A 28-year-old female with a history of gestational diabetes mellitus diagnosed eight years prior to presentation and subsequent type two diabetes mellitus ( T2DM ) , one prior episode of HTG-induced pancreatitis three years prior to presentation , associated with an acute hepatitis , and obesity with a body mass index ( BMI ) of 33.5 kg/m <sup>2</sup> , presented with a one-week history of polyuria , polydipsia , poor appetite , and vomiting .		0	433
1	Two weeks prior to presentation , she was treated with a five-day course of amoxicillin for a respiratory tract infection .		435	557
2	She was on metformin , glipizide , and dapagliflozin for T2DM and atorvastatin and gemfibrozil for HTG .		559	662
3	She had been on dapagliflozin for six months at the time of presentation .		664	737
4	Physical examination on presentation was significant for dry oral mucosa ; significantly , her abdominal examination was benign with no tenderness , guarding , or rigidity .		739	911
5	Pertinent laboratory findings on admission were : serum glucose 111 mg/dL , bicarbonate 18 mmol/L , anion gap 20 , creatinine 0.4 mg/dL , triglycerides 508 mg/dL , total cholesterol 122 mg/dL , glycated hemoglobin ( HbA1c ) 10% , and venous pH 7.27 .		913	1162
6	Serum lipase was normal at 43 U/L .		1164	1198
7	Serum acetone levels could not be assessed as blood samples kept hemolyzing due to significant lipemia .		1200	1303

document

sentence

```
sentenceDetector = SentenceDetectorDLModel\  
.pretrained("sentence_detector_dl_healthcare", "en","clinical/models")\  
.setInputCols(["document"])\  
.setOutputCol("sentence")
```

# Tokenizer / RegexTokenizer

- ✓ Splits words in a relevant format for NLP
- ✓ RegexTokenizer splits text by a regex pattern.
- ✓ Input for most of the DL models

	sentence	begin	end
0	A 28-year-old female with a history of gestational diabetes mellitus diagnosed eight years prior to presentation and subsequent type two diabetes mellitus ( T2DM ) , one prior episode of HTG-induced pancreatitis three years prior to presentation , associated with an acute hepatitis , and obesity with a body mass index ( BMI ) of 33.5 kg/m <sup>2</sup> , presented with a one-week history of polyuria , polydipsia , poor appetite , and vomiting .	0	433
1	Two weeks prior to presentation , she was treated with a five-day course of amoxicillin for a respiratory tract infection .	435	557
2	She was on metformin , glipizide , and dapagliflozin for T2DM and atorvastatin and gemfibrozil for HTG .	559	662
3	She had been on dapagliflozin for six months at the time of presentation .	664	737
4	Physical examination on presentation was significant for dry oral mucosa ; significantly , her abdominal examination was benign with no tenderness , guarding , or rigidity .	739	911
5	Pertinent laboratory findings on admission were : serum glucose 111 mg/dl , bicarbonate 18 mmol/l , anion gap 20 , creatinine 0.4 mg/dL , triglycerides 508 mg/dL , total cholesterol 122 mg/dL , glycated hemoglobin ( HbA1c ) 10% , and venous pH 7.27 .	913	1162
6	Serum lipase was normal at 43 U/L .	1164	1198
7	Serum acetone levels could not be assessed as blood samples kept hemolyzing due to significant lipemia .	1200	1303

sentence / document

```
tokenizer = Tokenizer()\n    .setInputCols(["sentence"])\n    .setOutputCol("token")
```



	token	begin	end	sentence
0	A	0	0	0
1	28-year-old	2	12	0
2	female	14	19	0
3	with	21	24	0
4	a	26	26	0
5	history	28	34	0
6	of	36	37	0
7	gestational	39	49	0
8	diabetes	51	58	0
9	mellitus	60	67	0
73	Two	435	437	1
74	weeks	439	443	1
75	prior	445	449	1
76	to	451	452	1
77	presentation	454	465	1
78	,	467	467	1
79	she	469	471	1
80	was	473	475	1
81	treated	477	483	1
82	with	485	488	1

token

```
regexTokenizer = RegexTokenizer() \\n    .setInputCols(["document"]) \\n    .setOutputCol("regexToken") \\n    .setToLowercase(True) \\n    .setPattern("\\s+")
```

# WordEmbeddingsModel

- ✓ Pretrained models can be loaded maps tokens to vectors
- ✓ `embeddings_clinical` can be preferred in clinical documents

	token	begin	end	sentence
0	A	0	0	0
1	28-year-old	2	12	0
2	female	14	19	0
3	with	21	24	0
4	a	26	26	0
5	history	28	34	0
6	of	36	37	0
7	gestational	39	49	0
8	diabetes	51	58	0
9	mellitus	60	67	0

	token	begin	end	sentence
73	Two	435	437	1
74	weeks	439	443	1
75	prior	445	449	1
76	to	451	452	1
77	presentation	454	465	1
78	,	467	467	1
79	she	469	471	1
80	was	473	475	1
81	treated	477	483	1
82	with	485	488	1



```
+-----+  
| | embeddings |  
+-----+  
| [-0.570580005645754 0.44183000922203064 0.7010200023651123 -0.417... |  
| [-0.542639970779419 0.41475999355316161 0.0321999788284302 -0.4024... |  
| [-0.2708599865436554 0.04400600120425224 -0.020260000601410866 -0... |  
| [0.6191999912261963 0.14650000631809235 -0.08592499792575836 -0.2... |  
| [-0.3397899866104126 0.20940999686717987 0.46347999572753906 -0.6... |  
+-----+
```

```
word_embeddings =  
WordEmbeddingsModel.pretrained("embeddings_clinical","en","clinical/models")\  
.setInputCols(["sentence","token"])\\  
.setOutputCol("embeddings")\
```

embeddings

# MedicalNerModel/NerModel

- ✓ Assigns an NER label to each token.
- ✓ The default model is `ner_clinical`, if no name is provided.

	token	begin	end
0	A	0	0
1	28-year-old	2	12
2	female	14	19
3	with	21	24
4	a	26	26
5	history	28	34
6	of	36	37
7	gestational	39	49
8	diabetes	51	58
9	mellitus	60	67



	token	begin	end	label	confidence
0	A	0	0	O	0.9998
1	28-year-old	2	12	O	0.9991
2	female	14	19	O	0.996
3	with	21	24	O	0.9998
4	a	26	26	O	0.9989
5	history	28	34	O	0.8485
6	of	36	37	O	0.9613
7	gestational	39	49	B-PROBLEM	0.8199
8	diabetes	51	58	I-PROBLEM	0.9809
9	mellitus	60	67	I-PROBLEM	0.9607

**token**

**named\_entity**

```
clinical_ner = MedicalNerModel.pretrained("ner_clinical_large","en","clinical/models")\
    .setInputCols(["sentence", "token", "embeddings"])\
    .setOutputCol("ner")\
    .setLabelCasing("upper")
```

# BertForTokenClassifier

- ✓ Loads Bert Models with a token classification head for Named-Entity-Recognition (NER) tasks.

“A 28-year-old female with a history of gestational diabetes mellitus diagnosed eight years prior to presentation and subsequent type two diabetes mellitus”

	begin	end	token	label	metadata
0	0	0	A	O	{'sentence': '0', 'Some(O)': '0.999487', 'Some(B-TEST)': '4.54522E-5', 'Some(I-PROBLEM...}
1	2	12	28-year-old	O	{'sentence': '0', 'Some(O)': '0.99780965', 'Some(B-TEST)': '2.268446E-4', 'Some(I-PROB...}
2	14	19	female	O	{'sentence': '0', 'Some(O)': '0.99946886', 'Some(B-TEST)': '2.7098458E-5', 'Some(I-PRO...}
3	21	24	with	O	{'sentence': '0', 'Some(O)': '0.99975836', 'Some(B-TEST)': '8.164151E-6', 'Some(I-PROB...}
4	26	26	a	O	{'sentence': '0', 'Some(O)': '0.99976975', 'Some(B-TEST)': '1.953544E-5', 'Some(I-PROB...}
5	28	34	history	O	{'sentence': '0', 'Some(O)': '0.99965894', 'Some(B-TEST)': '1.4398003E-5', 'Some(I-PRO...}
6	36	37	of	O	{'sentence': '0', 'Some(O)': '0.9997003', 'Some(B-TEST)': '6.0822117E-6', 'Some(I-PROB...}
7	39	49	gestational	B-PROBLEM	{'sentence': '0', 'Some(O)': '1.2003799E-5', 'Some(B-TEST)': '1.2508507E-5', 'Some(I-P...}
8	51	58	diabetes	I-PROBLEM	{'sentence': '0', 'Some(O)': '6.9626126E-6', 'Some(B-TEST)': '3.292549E-6', 'Some(I-PR...}
9	60	67	mellitus	I-PROBLEM	{'sentence': '0', 'Some(O)': '1.0308599E-5', 'Some(B-TEST)': '2.3517548E-6', 'Some(I-P...}
10	69	77	diagnosed	O	{'sentence': '0', 'Some(O)': '0.99924546', 'Some(B-TEST)': '1.260147E-5', 'Some(I-PROB...}

```
tokenClassifier = MedicalBertForTokenClassifier.pretrained("bert_token_classifier_ner_clinical", "en", "clinical/models")\
    .setInputCols("token", "sentence")\
    .setOutputCol("ner")\
    .setCaseSensitive(True)
```

# NerConverter / NerConverterInternal

- ✓ Converts a IOB or IOB2 representation of NER to a user-friendly one, by associating the tokens of recognized entities and their label

	label	begin	end	words	confidence
0	O	27	27	a	0.9989
1	O	29	35	history	0.8485
2	O	37	38	of	0.9613
3	B-PROBLEM	40	50	gestational	0.8199
4	I-PROBLEM	52	59	diabetes	0.9809
5	I-PROBLEM	61	68	mellitus	0.9607

named\_entity



	chunks	begin	end	entities	confidence
0	gestational diabetes mellitus	40	68	PROBLEM	0.9205
1	subsequent type two diabetes mellitus	118	154	PROBLEM	0.75560004
2	T2DM	158	161	PROBLEM	0.9928
3	HTG-induced pancreatitis	187	210	PROBLEM	0.97975004
4	an acute hepatitis	264	281	PROBLEM	0.9519333
5	obesity	289	295	PROBLEM	0.997

chunk

```
from sparknlp_jsl.annotator import NerConverterInternal
from sparknlp.annotator import NerConverter

ner_converter = NerConverterInternal()\
    .setInputCols(["sentence", "token", "ner"])\\
    .setOutputCol("ner_chunk")\

# Optional
    .setWhiteList(["PROBLEM"]) # List of entities to process
    .setBlackList(["TREATMENT", "TEST"]) # List of entities to be excluded.
    .setThreshold(0.99) # Confidence threshold to filter the chunk entities.
    .setReplaceLabels({"Drug_BrandName": "Drug"}) # Replace label in only NerConverterInternal
```

# ChunkMerger

- ✓ Merges chunk columns coming from two or more annotators.
- ✓ NER, ContextualParser, TextMatcher, or any other annotator producing chunks are supported.

"A 63 years old man presents to the hospital with a history of recurrent infections that include cellulitis, pneumonias, and upper respiratory tract infections."

The diagram illustrates the process of merging two NER chunks into one. On the left, there are two separate tables: 'deid\_ner\_chunk' and 'clinical\_ner\_chunk'. An arrow points from these two tables to a third table on the right, labeled 'merged\_ner\_chunk'. The 'deid\_ner\_chunk' table has columns: begin, end, chunk, entity. It contains one row with values 0, 3, 4, AGE. The 'clinical\_ner\_chunk' table has columns: begin, end, chunk, entity. It contains four rows with values: (0, 63, 82, recurrent infections PROBLEM), (1, 97, 106, cellulitis PROBLEM), (2, 109, 118, pneumonias PROBLEM), and (3, 126, 159, upper respiratory tract infections PROBLEM). The 'merged\_ner\_chunk' table has columns: begin, end, chunk, entity. It contains five rows, combining the data from both tables: (0, 3, 4, AGE), (1, 63, 82, recurrent infections PROBLEM), (2, 97, 106, cellulitis PROBLEM), (3, 109, 118, pneumonias PROBLEM), and (4, 126, 159, upper respiratory tract infections PROBLEM).

deid_ner_chunk				clinical_ner_chunk				merged_ner_chunk						
begin	end	chunk	entity	begin	end	chunk	entity	begin	end	chunk	entity			
0	3	4	63	AGE	0	63	82	recurrent infections	PROBLEM	0	3	4	63	AGE
					1	97	106	cellulitis	PROBLEM	1	63	82	recurrent infections	PROBLEM
					2	109	118	pneumonias	PROBLEM	2	97	106	cellulitis	PROBLEM
					3	126	159	upper respiratory tract infections	PROBLEM	3	109	118	pneumonias	PROBLEM
								4	126	159	upper respiratory tract infections	PROBLEM		

```
chunk_merger = ChunkMergeApproach()\
    .setInputCols('clinical_ner_chunk', "deid_ner_chunk")\
    .setOutputCol('merged_ner_chunk')\
#Optional
    .setMergeOverlapping()\ # Sets whether to merge overlapping matched chunks.
    .setFalsePositivesResource()\ # Sets file with false positive pairs
    .setReplaceDictResource()\ # Sets replace dictionary pairs
    .setChunkPrecedence()\ # Sets what is the precedence when two chunks have the same start and end indices.
    .setBlackList()\ # If defined, list of entities to ignore. The rest will be processed.
```

# ChunkMapper/ DocMapper

- ✓ Maps entities with their correspondings which are based on pre-defined dictionary
  - ✓ Very useful in medical coding resolution.
  - ✓ Supports multi dictionaries in a single model

"The patient was female and patient of Dr. X. and she was given Dermovate, Aspargin"

	ner_chunk	begin	end
0	Dermovate	63	71
1	Aspagan	74	80



	ner_chunk	begin	end	mapping_result	relation	all_relations
0	Dermovate	63	71	lupus	treatment	discoid lupus erythematosus:::emepeines:::psori...
1	Aspagin	74	80	ankylosing spondylitis	treatment	arthralgia:::pain:::bursitis:::headache:::migr...

## token

## named\_entity

```
chunkerMapper_treatment= ChunkMapperModel().pretrained("drug_action_treatment_mapper", "en", "clinical/models")\
    .setInputCols(["ner_chunk"])\\
    .setOutputCol("mappings")\
    .setRels(["treatment"])# should be one of the relations from the model
```

# DrugNormalizer

- ✓ Transforms text to the format used in the RxNorm and SNOMED standards

	text		normalized_text
0	Sodium Chloride/Potassium Chloride 13bag		Sodium Chloride / Potassium Chloride 13 bag
1	interferon alfa-2b 10 million unit ( 1 ml ) injec		interferon alfa - 2b 10000000 unt ( 1 ml ) injection
2	aspirin 10 meq/ 5 ml oral sol		aspirin 2 meq/ml oral solution

```
drugNormalizer = DrugNormalizer()\
    .setInputCols(["document"])\
    .setOutputCol("document_normalized")
#Optional
.setPolicy()\ # all / abbreviations / dosages
```

# ContextualParser

- ✓ Allows users to extract entities from a document based on pattern matching.

A  28 year old female with a history of gestational diabetes mellitus diagnosed 8 years ago.

AGE

3 years ago, he reported an episode of HTG-induced pancreatitis .

5

months old boy with repeated concussions.

AGE

```
age = {  
    "entity": "Age",  
    "ruleScope": "sentence",  
    "matchScope": "token",  
    "regex": "\d{1,3}",  
    "prefix": ["age of", "age"],  
    "suffix": ["-years-old", "years-old", "-year-old", "-months-old",  
              "-month-old", "-months-old", "-day-old", "years old",  
              "-days-old", "month old", "years", "year", "months", "old",  
              "days old", "year old"],  
    "contextLength": 25,  
    "contextException": ["ago"],  
    "exceptionDistance": 12}
```

```
age_contextual_parser = ContextualParserApproach()\\"  
    .setInputCols(["sentence", "token"]) \  
    .setOutputCol("chunk_age") \  
    .setJsonPath("age.json") \  
    .setCaseSensitive(False) \  
    .setPrefixAndSuffixMatch(False)\ \  
    .setShortestContextMatch(True)\ \  
    .setOptionalContextRules(False)
```

# Deidentification

- ✓ Allows users to mask or replace PHI (Protected Health Information) entities.

**sentence**

```
Patient AIQING, 25 years-old (ssid : 321-55-3699), born in Beijing,  
was transferred to the The Johns Hopkins Hospital.
```



	<b>deid_entity_label</b>
	Patient <NAME>, <AGE> years-old (ssid : <ID>), born in <LOCATION>, was transferred to the The <LOCATION>
	<b>deid_same_length</b>
	Patient [***], ** years-old (ssid : [*****]), born in [***], was transferred to the The [*****]
	<b>deid_fixed_length</b>
	Patient *****; ***** years-old (ssid : *****), born in *****; ***** was transferred to the The *****
	<b>obfuscated</b>
	Patient Olene Ronde, 41 years-old (ssid : G074713), born in China, was transferred to the Avenue Hospital

```
deidentification = DeIdentification()\n    .setInputCols(["sentence", "token", "ner_chunk"])\n    .setOutputCol("deidentified")\n    .setMode("mask") # "obfuscate"\n    # Optional\n    .setReturnEntityMappings(True)\n    .setMaskingPolicy("entity_labels") # "same_length_chars", "deid_fixed_length", "fixed_length_chars"\n    .setObfuscateDate(False)\n    .setObfuscateRefSource("file") # "file", "both"
```

# Spell Checker

- ✓ A flexible, configurable and "re-usable by parts" model.
- ✓ Different correction candidates for each word - **word level**.
- ✓ The surrounding text of each word, i.e. its context - **sentence level**.
- ✓ The relative cost of different correction candidates according to the edit operations at the character level it requires - **subword level**.

"Alliow me tao introdouce myhelf, I am  
a man of waelth und tiaste"



"Allow me to introduce myself, I am a  
man of wealth and taste"

```
spellModel = ContextSpellCheckerModel\  
.pretrained('spellcheck_dl')\  
.setInputCols("token")\  
.setOutputCol("checked")\  
.setErrorThreshold(4.0) \  
.setTradeoff(6.0)
```

# TextMatcher

- ✓ Matches exact phrases (by token) provided in a file against a Document.
- ✓ A text file of predefined phrases must be provided with `setEntities`

"A 28-year-old female with a history  
of gestational diabetes mellitus and  
coronary artery disease"



	begin	end	entity	entity
0	39	67	gestational diabetes mellitus	DISEASE
1	73	95	coronary artery disease	DISEASE

```
entityExtractor = TextMatcher() \
    .setInputCols(["document", "token"]) \
    .setEntities("phrases.txt", ReadAs.TEXT) \
    .setOutputCol("entity") \
    .setEntityValue("DISEASE") \
    .setCaseSensitive(False)
```

Phrases.txt

-----

gestational diabetes mellitus

coronary artery disease

Lower respiratory infections

# Normalizer

- ✓ Removes all dirty characters from text following a regex pattern and transforms words based on a provided dictionary

"#publiHUCA #Emergency Emergency department observation of patients with acute heart failure prior to hospital admission: impact on short-term prognosis "



token	token	normalized
0 #publiHUCA	0 #publiHUCA	publihuca
1 #Emergency	1 #Emergency	emergency
2 Emergency	2 Emergency	emergency
3 department	3 department	department
4 observation	4 observation	observation

```
normalizer = Normalizer() \
    .setInputCols(["token"]) \
    .setOutputCol("normalized") \
    .setLowercase(True) \ # lowercase tokens
    .setCleanupPatterns(["[^\\w\\d\\s]"]) # remove punctuations (keep alphanumeric chars)
    .setSlangDictionary(path) # txt file with delimited words to be transformed into something else
```

# DocumentNormalizer

- ✓ Normalize documents once that they have been processed and indexed

```
<div id="theworldsgreatest" class='my-right my-hide-small my-wide toptext'  
style="font-family:'Segoe UI',Arial,sans-serif">  
    THE WORLD'S LARGEST WEB DEVELOPER SITE  
    <h1 style="font-size:300%;">THE WORLD'S LARGEST WEB DEVELOPER SITE</h1>  
    <p style="font-size:160%;">Lorem Ipsum is simply dummy text of the printing  
and typesetting industry. Lorem Ipsum has been the industry's standard dummy  
text ever since the 1500s, when an unknown printer took a galley of type and  
scrambled it to make a type specimen book.</p>  
</div> </div> "
```

document



"the world's largest web developer site the world's  
largest web developer site lorem ipsum is simply dummy  
text of the printing and typesetting industry. lorem  
ipsum has been the industry's standard dummy text ever  
since the 1500s, when an unknown printer took a galley of  
type and scrambled it to make a type specimen book. "

document

```
documentNormalizer = DocumentNormalizer() \  
.setInputCols("document") \  
.setOutputCol("normalizedDocument") \  
.setAction("clean") \  
.setPatterns(cleanUpPatterns) \  
.setReplacement(" ") \  
.setPolicy("pretty_all") \  
.setLowercase(True)
```

# RegexMatcher

- ✓ Uses a reference file to match a set of regular expressions and associate them with a provided identifier.

'The patient (123-45-7890) was  
discharged from the hospital on 10th  
January.'



	clinical_entities	begin	end	label
0	123-45-7890	13	23	SSN

```
regex_matcher = RegexMatcher()\n    .setInputCols('document')\n    .setStrategy("MATCH_ALL")\n    .setOutputCol("regex_matches")\n    .setExternalRules(path='./regex_rules.txt', delimiter=',')
```

```
regex_rules.txt\n-----\n\d{3}.\?\d{2}.\?\d{4}, SSN
```

# NerOverwriter

- ✓ Overwrites or replaces the labels of entities of specified strings.

"Mr.Brown was described warfarin 5 mg"

token	ner_label
0 Mr.Brown	O
1 was	O
2 described	O
3 warfarin	B-TREATMENT
4 5	O
5 mg	O



token	ner_label	ner_overwritten
0 Mr.Brown	O	B-People
1 was	O	O
2 described	O	O
3 warfarin	B-TREATMENT	B-DRUG
4 5	O	O
5 mg	O	O

```
nerOverwriter = NerOverwriter() \
    .setInputCols(["ner"]) \
    .setOutputCol("ner_overwritten") \
    .setNerWords(["Mr.Brown"]) \
    .setNewNerEntity("B-People") \
    .setReplaceEntities({"B-TREATMENT" : "B-DRUG"})
```

# BertSentenceEmbeddings

- ✓ Calculates BERT embeddings for sequence

"Patient has a headache for the last 2 weeks, needs to get a head CT, and appears anxious when she walks fast. No alopecia noted. She denies pain"

	sentence	sentence_embeddings
0	Patient has a headache for the last 2 weeks, n...	-0.08882622 -0.09253775 -0.025744002 -0.105720...

```
Classifier = BertForSequenceClassification.pretrained("bert_sequence_classifier_covid_sentiment", "en", "clinical/models")\
.setInputCols(["document", "token"])\\
.setOutputCol("class")
```

# BertForSequenceClassifier

- ✓ Loads BERT Models with sequence classification/regression for multi-class document classification tasks.

		text	result
0		British Department of Health confirms first two cases of in UK	[neutral]
1	so my trip to visit my australian exchange student just got canceled bc of coronavirus. im heartbroken :(		[negative]
2	I wish everyone to be safe at home and stop pandemic		[positive]

```
Classifier = BertForSequenceClassification.pretrained("bert_sequence_classifier_covid_sentiment", "en", "clinical/models")\
    .setInputCols(["document", "token"])\\
    .setOutputCol("class")
```

# AssertionDLModel

- ✓ A Deep Learning based approach is used to extract Assertion Status from extracted entities and text.

"Patient has a headache for the last 2 weeks, needs to get a head CT, and appears anxious when she walks fast. No alopecia noted. She denies pain"

	entities_ner_chunk	entities_ner_chunk_class	assertion
0	a headache	PROBLEM	present
0	a head CT	TEST	present
0	anxious	PROBLEM	present
0	alopecia	PROBLEM	absent
0	pain	PROBLEM	absent

```
clinical_assertion = AssertionDLModel.pretrained("assertion_dl", "en", "clinical/models") \
    .setInputCols(["sentence", "ner_chunk", "embeddings"]) \
    .setOutputCol("assertion")
```

# Relation Extraction Model

- ✓ Extracts and classifies instances of relations between named entities.
- ✓ Relation pairs need to be defined with setRelationPairs, to specify between which entities the extraction should be done.

"I experienced fatigue, muscle cramps, anxiety, aggression and sadness after taking Lipitor but no more adverse after passing Zocor."



chunk1	entity1	entity2_begin	entity2_end	chunk2	entity2	relation	score
fatigue	ADE	82	88	Lipitor	DRUG	1	
fatigue	ADE	124	128	Zocor	DRUG	0	
muscle cramps	ADE	82	88	Lipitor	DRUG	1	
muscle cramps	ADE	124	128	Zocor	DRUG	0	
anxiety	ADE	82	88	Lipitor	DRUG	1	
anxiety	ADE	124	128	Zocor	DRUG	0	
aggression	ADE	82	88	Lipitor	DRUG	1	
aggression	ADE	124	128	Zocor	DRUG	0	
sadness	ADE	82	88	Lipitor	DRUG	1	
sadness	ADE	124	128	Zocor	DRUG	0	

```
reModel = RelationExtractionModel()\
    .pretrained("re_ade_clinical", "en", 'clinical/models')\
    .setInputCols(["embeddings", "pos_tags", "ner_chunks", "dependencies"])\
    .setOutputCol("relations")\
    .setMaxSyntacticDistance(10) \
    .setRelationPairs(["drug-ade, ade-drug"])\
    .setRelationPairsCaseSensitive(False)
```

# Relation Extraction DL Model

- ✓ A Deep Learning based approach is used to extract Assertion Status from extracted entities and text.

“The patient was prescribed 1 unit of Advil for 5 days after meals. The patient was also given 1 unit of Metformin daily. He was seen by the endocrinology service and she was discharged on 40 units of insulin glargine at night.”

sentence	entity1_begin	entity1_end	chunk1	entity1	entity2_begin	entity2_end	chunk2	entity2	relation	
0	0	28	33	1 unit	DOSAGE	38	42	Advil	DRUG	DOSAGE-DRUG
1	0	38	42	Advil	DRUG	44	53	for 5 days	DURATION	DRUG-DURATION
2	1	96	101	1 unit	DOSAGE	106	114	Metformin	DRUG	DOSAGE-DRUG
3	1	106	114	Metformin	DRUG	116	120	daily	FREQUENCY	DRUG-FREQUENCY
4	2	190	197	40 units	DOSAGE	202	217	insulin glargine	DRUG	DOSAGE-DRUG
5	2	202	217	insulin glargine	DRUG	219	226	at night	FREQUENCY	DRUG-FREQUENCY

```
reModel = RelationExtractionModel().pretrained("posology_re")\
    .setInputCols(["embeddings", "pos_tags", "ner_chunks", "dependencies"])\
    .setOutputCol("relations")\
    .setMaxSyntacticDistance(4)
```

# Pretrained NER Models

- Clinical NER Models

index	model_name	index	model_name	index	model_name	index	model_name
1	jsl_ner_wip_clinical	17	ner_chexpert	33	ner_deid_subentity (German)	49	ner_jsl_greedy
2	jsl_ner_wip_greedy_clinical	18	ner_clinical	34	ner_diseases_large	50	ner_jsl_slim
3	jsl_ner_wip_modifier_clinical	19	ner_clinical_icdem	35	ner_drugs	51	ner_measurements_clinical
4	jsl_rd_ner_wip_greedy_clinical	20	ner_clinical_large	36	ner_drugs_greedy	52	ner_medmentions_coarse
5	ner_ade_clinical	21	ner_clinical_large_en	37	ner_drugs_large	53	ner_posology
6	ner_ade_clinicalalbert	22	ner_deid_augmented	38	ner_events_admission_clinical	54	ner_posology_experimental
7	ner_ade_healthcare	23	ner_deid_enriched	39	ner_events_clinical	55	ner_posology_greedy
8	ner_anatomy	24	ner_deid_generic_augmented	40	ner_events_healthcare	56	ner_posology_healthcare
9	ner_anatomy_coarse	25	ner_deid_generic (German)	41	ner_genetic_variants	57	ner_posology_large
10	ner_bacterial_species	26	ner_deid_large	42	ner_healthcare	58	ner_posology_small
11	ner_bionlp	27	ner_deid_sd	43	ner_human_phenotype_gene_clinical	59	ner_profiling_clinical
12	ner_cancer_genetics	28	ner_deid_sd_large	44	ner_human_phenotype_go_clinical	60	ner_radiology
13	ner_cellular	29	ner_deid_subentity_augmented	45	ner_jsl	61	ner_radiology_wip_clinical
14	ner_chemicals	30	ner_deid_synthetic	46	ner_jsl_enriched	62	ner_risk_factors
15	ner_chemprot_clinical	31	ner_deidentify_dl	47	ner_nihss	63	ner biomarker
16	ner_abbreviation_clinical	32	ner_deid_subentity_augmented_i2b2	48	ner_diseases	64	ner_drugprot_clinical

- BioBert NER Models

index	model_name	index	model_name	index	model_name	index	model_name
1	jsl_ner_wip_greedy_biobert	7	ner_cellular_biobert	13	ner_events_biobert	18	ner_jsl_greedy_biobert
2	jsl_rd_ner_wip_greedy_biobert	8	ner_chemprot_biobert	14	ner_human_phenotype_gene_biobert	19	ner_posology_biobert
3	ner_ade_biobert	9	ner_clinical_biobert	15	ner_human_phenotype_go_biobert	20	ner_posology_large_biobert
4	ner_anatomy_biobert	10	ner_deid_biobert	16	ner_jsl_biobert	21	ner_profiling_biobert
5	ner_anatomy_coarse_biobert	11	ner_deid_enriched_biobert	17	ner_jsl_enriched_biobert	22	ner_risk_factors_biobert
6	ner_bionlp_biobert	12	ner_diseases_biobert				

- BertForTokenClassification Clinical NER models

model_name
1 bert_token_classifier_ner_ade
2 bert_token_classifier_ner_clinical
3 bert_token_classifier_ner_deid
4 bert_token_classifier_ner_drugs
5 bert_token_classifier_ner_jsl
6 bert_token_classifier_ner_jsl_slim
7 bert_token_classifier_ner_bionlp
8 bert_token_classifier_ner_bacteria
9 bert_token_classifier_ner_anatomy
10 bert_token_classifier_ner_cellular
11 bert_token_classifier_ner_chemprot
12 bert_token_classifier_ner_chemicals
13 bert_token_classifier_drug_development_trials

Approach	embeddings	# of models
BiLSTM-CNN-Char	Clinical (glove)	100+
BiLSTM-CNN-Char	Biobert	30+
Bert for Token Cls.	Biobert	20+
Total		150+

# NER JSL

Let's show an example of `ner_js1` model that has about 80 clinical entity labels by changing just only the model name.

## Entities

Injury_or_Poisoning	Direction	Test	Admission_Discharge	Death_Entity
Relationship_Status	Duration	Respiration	Hyperlipidemia	Birth_Entity
Age	Labour_Delivery	Family_History_Header	BMI	Temperature
Alcohol	Kidney_Disease	Oncological	Medical_History_Header	Cerebrovascular_Disease
Oxygen_Therapy	O2_Saturation	Psychological_Condition	Heart_Disease	Employment
Obesity	Disease_Syndrome_Disorder	Pregnancy	ImagingFindings	Procedure
Medical_Device	Race_Ethnicity	Section_Header	Symptom	Treatment
Substance	Route	Drug_Ingredient	Blood_Pressure	Diet
External_body_part_or_region	LDL	VS_Finding	Allergen	EKG_Findings
Imaging_Technique	Triglycerides	RelativeTime	Gender	Pulse
Social_History_Header	Substance_Quantity	Diabetes	Modifier	Internal_organ_or_component
Clinical_Dept	Form	Drug_BrandName	Strength	Fetus_NewBorn
RelativeDate	Height	Test_Result	Sexually_Active_or_Sexual_Orientation	Frequency
Time	Weight	Vaccine	Vital_Signs_Header	Communicable_Disease
Dosage	Overweight	Hypertension	HDL	Total_Cholesterol
Smoking	Date			

# Attention (aka Bert) is all you need ?

ner model	embeddings_clinical (BLSTM-CNN-Char)		biobert (BLSTM-CNN-Char)		BertForTokenClassification (SOTA)	
	micro	macro	micro	macro	micro	macro
ner_jsl	0.878	<b>0.814</b>	0.862	0.711	<b>0.88</b>	0.71
ner_jsl_slim	0.87	0.766	0.86	<b>0.778</b>	<b>0.89</b>	0.75
ner_deid	<b>0.94</b>	<b>0.77</b>	0.93	0.77	0.75	0.63
ner_drug	0.964	0.964	0.912	0.911	<b>1</b>	<b>0.98</b>
ner_ade	0.84	0.807	0.839	0.819	<b>0.89</b>	<b>0.84</b>

\* On average, the GLoVe embeddings are 30% faster during training compared to BERT embeddings, and more than 5x faster during inference, while being on-par in terms of F1 score.

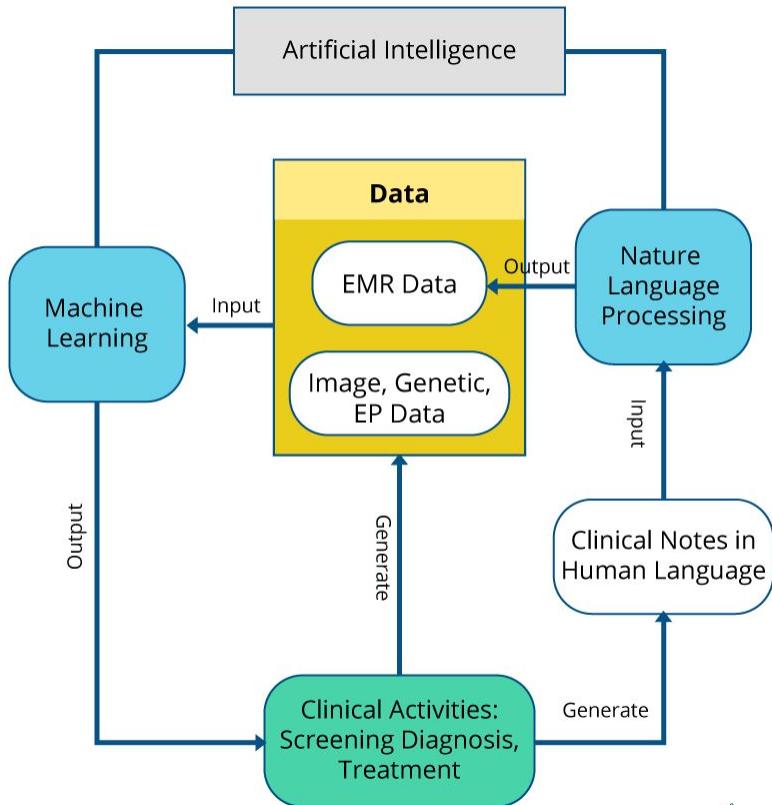
# NLP in Healthcare

"Mother with a lung cancer, a patient is diagnosed as breast cancer in 1991 and then admitted to Mayo Clinic in Oct 2000, went under chemo for 6 months, discharged in April 2001 with a prescription of 2 mg metformin 3 times per day."

## Named Entities

Mother with a lung cancer **ONCOLOGICAL** , a pregnant **PREGNANCY** patient is diagnosed as breast cancer **ONCOLOGICAL** in **1991 DATE** and then admitted **ADMISSION\_DISCHARGE** to Mayo Clinic **CLINICAL\_DEPT** in Oct **2000 DATE** , went under chemo **TREATMENT** for 6 months **DURATION** , discharged **ADMISSION\_DISCHARGE** in **April 2001 DATE** with a prescription of **2 mg STRENGTH** metformin **DRUG\_INGREDIENT** **3 times per day FREQUENCY** .

# Clinical Named Entity Recognition (NER)



The patient was prescribed 1 capsule of Advil for 5 days . He was seen by the endocrinology service and she was discharged on 40 units of insulin glargine at night , 12 units of insulin lispro with meals , and metformin 1000 mg two times a day . It was determined that all SGLT2 inhibitors should be discontinued indefinitely fro 3 months .

Color codes:FREQUENCY, DOSAGE, DURATION, DRUG, FORM, STRENGTH, **Posology NER**

No findings in urinary system , skin color is normal , brain CT and cranial checks are clear . Swollen fingers and eyes . Extensive stage small cell lung cancer . Chemotherapy with carboplatin and etoposide . Left scapular pain status post CT scan of the thorax .

Color codes:Organ, Organism\_subdivision, Organism\_substance, PathologicalFormation, Anatomical\_system, **Anatomy NER**

A . Record date : 2093-01-13 , David Hale , M.D . , Name : Hendrickson , Ora MR . # 7194334 Date : 01/13/93 PCP : Oliveira , 25 years-old , Record date : 2079-11-09 . Cocke County Baptist Hospital . 0295 Keats Street

Color codes:STREET, DOCTOR, AGE, HOSPITAL, PATIENT, DATE, MEDICALRECORD, **PHI NER**

# Clinical Named Entity Recognition (NER)

A 28-year-old female with a history of gestational diabetes mellitus diagnosed eight years prior to presentation and subsequent type two diabetes mellitus (T2DM), one prior episode of HTG-induced pancreatitis three years prior to presentation, and associated with an acute hepatitis, presented with a one-week history of polyuria, poor appetite, and vomiting. She was on metformin, glipizide, and dapagliflozin for T2DM and atorvastatin and gemfibrozil for HTG. She had been on dapagliflozin for six months at the time of presentation. Physical examination on presentation was significant for dry oral mucosa ; significantly , her abdominal examination was benign with no tenderness, guarding, or rigidity. Pertinent laboratory findings on admission were: serum glucose 111 mg/dL, creatinine 0.4 mg/dL, triglycerides 508 mg/dL, total cholesterol 122 mg/dL, and venous pH 7.27.

D

A 28-year-old female with a history of gestational diabetes mellitus diagnosed eight years prior to presentation and subsequent type two diabetes mellitus (T2DM), one prior episode of HTG-induced pancreatitis three years prior to presentation, and associated with an acute hepatitis, presented with a one-week history of polyuria, poor appetite, and vomiting. She was on metformin, glipizide, and dapagliflozin for T2DM and atorvastatin and gemfibrozil for HTG. She had been on dapagliflozin for six months at the time of presentation. Physical examination on presentation was significant for dry oral mucosa ; significantly , her abdominal examination was benign with no tenderness, guarding, or rigidity. Pertinent laboratory findings on admission were: serum glucose 111 mg/dL, creatinine 0.4 mg/dL, triglycerides 508 mg/dL, total cholesterol 122 mg/dL, and venous pH 7.27.

ner\_clinical

A 28-year-old female with a history of gestational diabetes mellitus diagnosed eight years prior to presentation and subsequent type two diabetes mellitus (T2DM), one prior episode of HTG-induced pancreatitis three years prior to presentation, and associated with an acute hepatitis, presented with a one-week history of polyuria, poor appetite, and vomiting. She was on metformin, glipizide, and dapagliflozin for T2DM and atorvastatin and gemfibrozil for HTG. She had been on dapagliflozin for six months at the time of presentation. Physical examination on presentation was significant for dry oral mucosa ; significantly , her abdominal examination was benign with no tenderness, guarding, or rigidity. Pertinent laboratory findings on admission were: serum glucose 111 mg/dL, creatinine 0.4 mg/dL, triglycerides 508 mg/dL, total cholesterol 122 mg/dL, and venous pH 7.27.

ner\_jsl

# Clinical Named Entities – Spark NLP vs Others

## Spark NLP

Google	Azure	AWS
PROBLEM	DIAGNOSIS SYMPTOM_OR_SIGN ALLERGEN	MEDICAL_CONDITION_DX_NA MEDICAL_CONDITION_SIGN MEDICAL_CONDITION_SYMPTOM
PROCEDURE	TREATMENT_NAME	PROCEDURE_NAME TREATMENT_NAME
MEDICINE	MEDICATION_CLASS MEDICATION_NAME	MEDICATION_BRAND_NAME MEDICATION_GENERIC_NAME
ANATOMICAL_STRUCTURE	BODY_STRUCTURE	SYSTEM_ORGAN_SITE
LABORATORY_DATA BODY_MEASUREMENT	EXAMINATION_NAME	TEST_NAME
SEVERITY	CONDITION_QUALIFIER CONDITION_SCALE	MEDICAL_CONDITION_ACUITY
MED_DOSE MED_TOTALDOSE MED_STRENGTH MED_UNIT	DOSAGE	MEDICATION_DOSAGE MEDICATION_STRENGTH MEDICATION_RATE
MED_FREQUENCY	FREQUENCY	MEDICATION_FREQUENCY
MED_FORM	MEDICATION_FORM	MEDICATION_FORM
MED_ROUTE	MEDICATION_ROUTE	MEDICATION_ROUTE_OR_MODE
MED_DURATION	TIME	MEDICATION_DURATION
LAB_VALUE MED_VALUE	MEASUREMENT_VALUE	TEST_VALUE
LAB_UNIT BM_UNIT	MEASUREMENT_UNIT	TEST_UNIT



Symptom, Disease\_Syndrome\_Disorder, Heart\_Disease, VS\_Finding, Communicable\_Disease, Hypertension, Diabetes, Kidney\_Disease, Cerebrovascular\_Disease, Injury\_or\_Poisoning, Psychological\_Condition, Total\_Cholesterol, Hyperlipidemia, Obesity, Oncological, Pregnancy, EKG\_Findings, Death\_Entity, ImagingFindings, Female\_Reproductive\_Status, Fetus\_NewBorn, Pregnancy\_Delivery\_Puerperium, Overweight, Puerperium



Test, Test\_Result, Treatment, Pulse, Imaging\_Technique, Labour\_Delivery, Temperature, Blood\_Pressure, Oxygen\_Therapy, Weight, LDL, O2\_Saturation, BMI, Vaccine, Respiration, Triglycerides

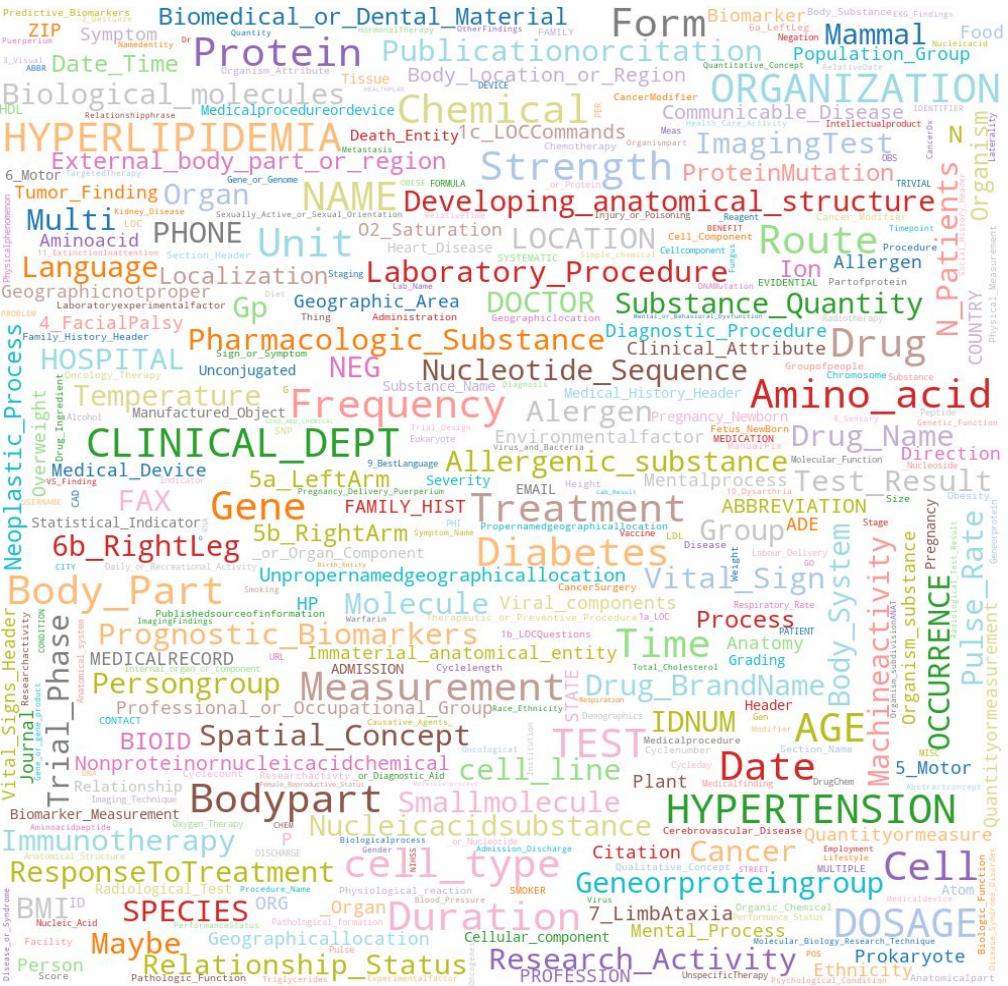
Mother with a lung cancer ONCOLOGICAL , a pregnant PREGNANCY patient is diagnosed as breast cancer ONCOLOGICAL in 1991 DATE and then admitted ADMISSION\_DISCHARGE to Mayo Clinic CLINICAL\_DEPT in Oct 2000 DATE , went under chemo TREATMENT for 6 months DURATION , discharged ADMISSION\_DISCHARGE in April 2001 DATE with a prescription of 2 mg STRENGTH metformin DRUG\_INGREDIENT 3 times per day FREQUENCY .

# Clinical Named Entities

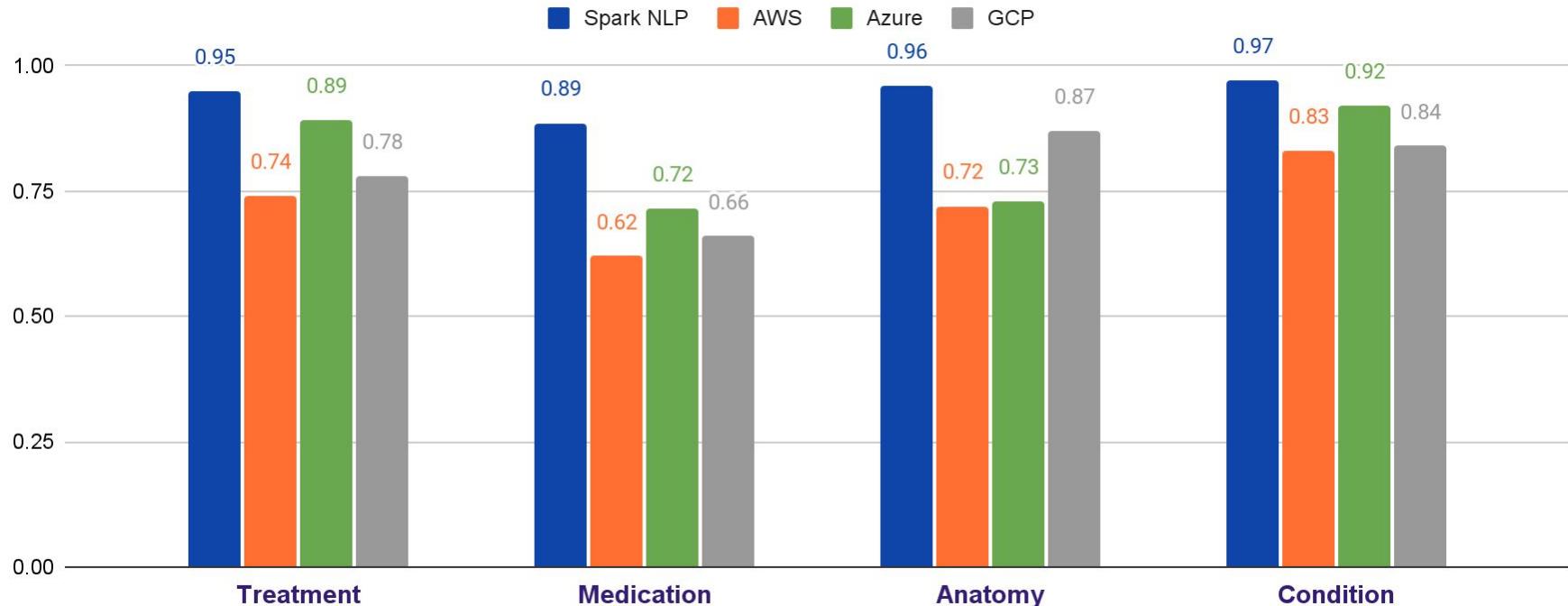
## Spark NLP vs Others

Google	Azure	AWS
PROBLEM	DIAGNOSIS SYMPTOM_OR_SIGN ALLERGEN	MEDICAL_CONDITION_DX_NA MEDICAL_CONDITION_SIGN MEDICAL_CONDITION_SYMPTOM
PROCEDURE	TREATMENT_NAME	PROCEDURE_NAME TREATMENT_NAME
MEDICINE	MEDICATION_CLASS MEDICATION_NAME	MEDICATION_BRAND_NAME MEDICATION_GENERIC_NAME
ANATOMICAL_STRUCTURE	BODY_STRUCTURE	SYSTEM_ORGAN_SITE
LABORATORY_DATA BODY_MEASUREMENT	EXAMINATION_NAME	TEST_NAME
SEVERITY	CONDITION_QUALIFIER CONDITION_SCALE	MEDICAL_CONDITION_ACUITY
MED_DOSE MED_TOTALDOSE MED_STRENGTH MED_UNIT	DOSAGE	MEDICATION_DOSAGE MEDICATION_STRENGTH MEDICATION_RATE
MED_FREQUENCY	FREQUENCY	MEDICATION_FREQUENCY
MED_FORM	MEDICATION_FORM	MEDICATION_FORM
MED_ROUTE	MEDICATION_ROUTE	MEDICATION_ROUTE_OR_MODE
MED_DURATION	TIME	MEDICATION_DURATION
LAB_VALUE MED_VALUE	MEASUREMENT_VALUE	TEST_VALUE
LAB_UNIT BM_UNIT	MEASUREMENT_UNIT	TEST_UNIT

## 400+ entities from 100+ models



# NER Benchmarks



# Spark NLP vs AWS vs GCP vs Academic

		Spark NLP	Competition Best	Last Best
Clinical Concept Extraction	2010 i2b2/VA	<b>0.876</b>	0.852	0.862
De-Identification	2014 n2c2	<b>0.961</b>	0.936	0.955
Medication Extraction	2018 n2c2	<b>0.899</b>	0.896	0.896

Entity	Sample	Spark NLP Clinical Models			AWS Medical Comprehend			GCP Healthcare API		
		Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
Problem	4891	0.726	0.585	<b>0.648</b>	0.539	0.478	<b>0.507</b>	0.850	0.516	0.642
Test	5903	0.782	0.662	<b>0.717</b>	0.594	0.703	<b>0.644</b>	0.576	0.461	0.512
Drug	10284	0.946	0.882	0.913	0.815	0.910	<b>0.860</b>	0.962	0.885	<b>0.922</b>
Avg. F1				<b>0.759</b>			<b>0.670</b>			0.692

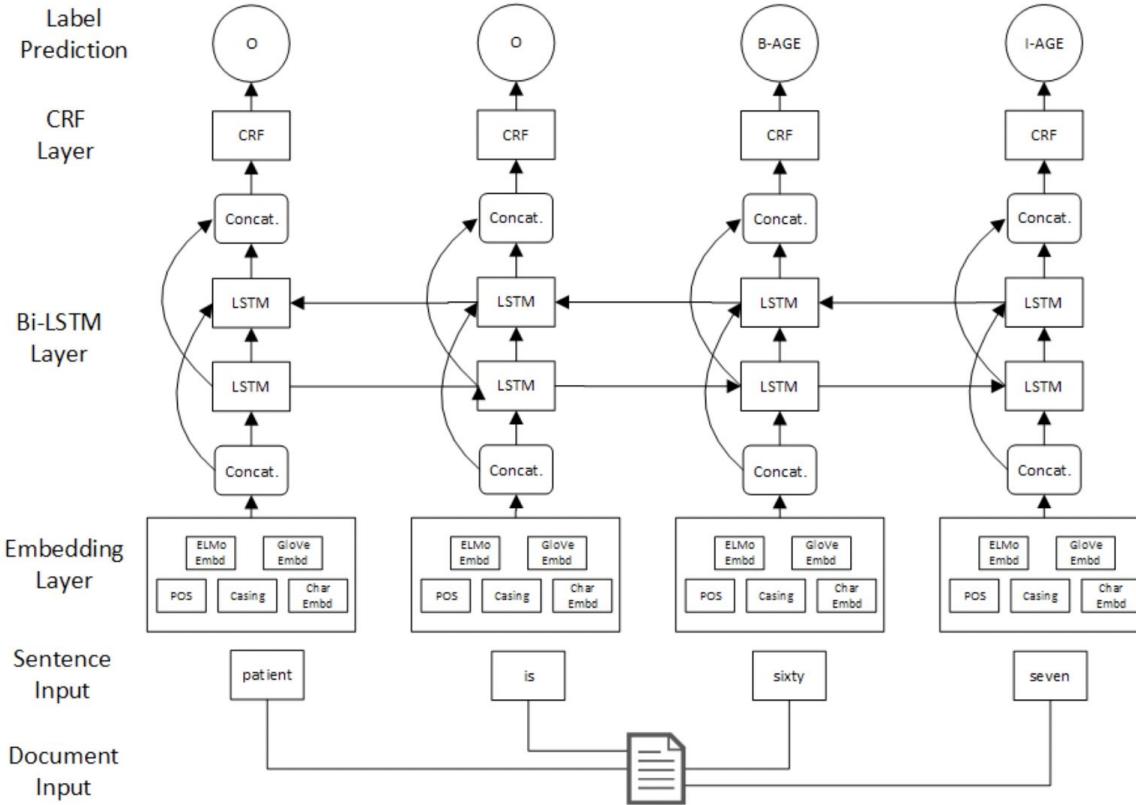
# Biomedical Named Entity Recognition

## Spark NLP vs Spacy vs Stanza

Dataset	Entities	Spark - Biomedical	Spark - GloVe 6B	Stanza	SciSpacy
NBCI-Disease	Disease	<b>89.13</b>	87.19	87.49	81.65
BC5CDR	Chemical, Disease	<b>89.73</b>	88.32	88.08	83.92
BC4CHEMD	Chemical	<b>93.72</b>	92.32	89.65	84.55
Linnaeus	Species	86.26	85.51	<b>88.27</b>	81.74
Species800	Species	<b>80.91</b>	79.22	76.35	74.06
JNLPBA	5 types in cellular	<b>81.29</b>	79.78	76.09	73.21
AnatEM	Anatomy	<b>89.13</b>	87.74	88.18	84.14
BioNLP13-CG	16 types in Cancer Genetics	<b>85.58</b>	84.3	84.34	77.6

Benchmarks on BioMedical NER Datasets

# NER Architecture



Char-CNN-BiLSTM

John	B-PER
Smith	I-PER
lives	0
in	0
New	B-LOC
York	I-LOC

John Smith  $\Rightarrow$  PERSON  
 New York  $\Rightarrow$  LOCATION

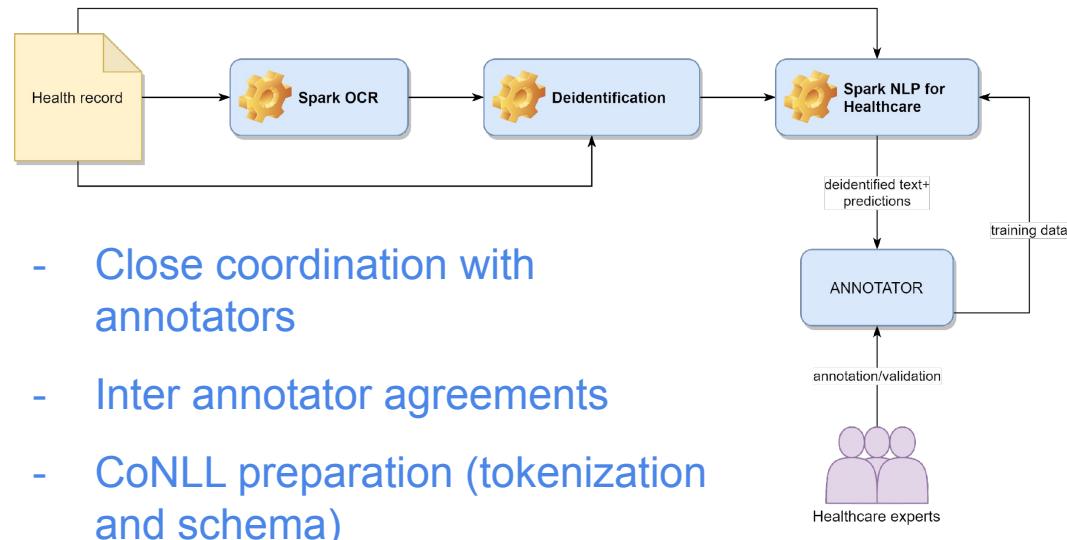
word	POS_tag	chunk_tag	NER_tag
She	PRP	O	B-person
presented	VBD	B-VP	O
with	IN	B-VP	O
left	JJ	B-NP	B-problem
upper	JJ	I-NP	I-problem
quadrant	NN	I-NP	I-problem
pain	NN	I-NP	I-problem
as	RB	O	O
well	RB	O	O
as	IN	B-VP	O
nausea	NN	B-NP	B-problem

John	B-PER
Smith	I-PER
lives	O
in	O
New	B-LOC
York	I-LOC

John Smith  $\Rightarrow$  PERSON  
 New York  $\Rightarrow$  LOCATION

word	POS_tag	chunk_tag	NER_tag
She	PRP	O	B-person
presented	VBD	B-VP	O
with	IN	B-VP	O
left	JJ	B-NP	B-problem
upper	JJ	I-NP	I-problem
quadrant	NN	I-NP	I-problem
pain	NN	I-NP	I-problem
as	RB	O	O
well	RB	O	O
as	IN	B-VP	O
nausea	NN	B-NP	B-problem

# NER in Healthcare



- Close coordination with annotators
- Inter annotator agreements
- CoNLL preparation (tokenization and schema)

She returns today for ongoing evaluation of her EGFR mutated, stage 4 lung cancer with metastasis to her L2 vertebrae and her lungs bilaterally.

Bone negative for metastatic disease.

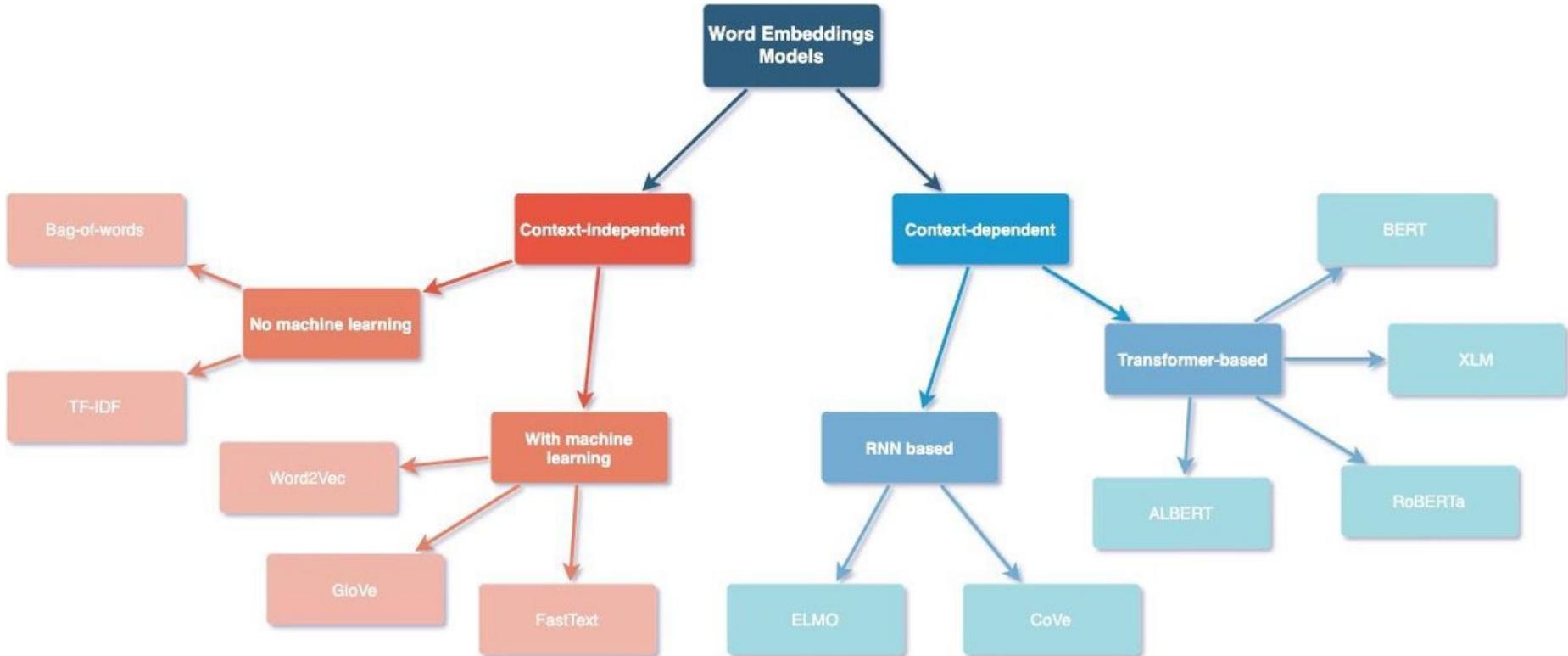
Patient denies any family history of cancer.

# NER-DL in Spark NLP

## Char-CNN-BiLSTM

	F1 : Tokens	F2 : Casing	F3 : POS	F4 : Char CNN	Labels
The					O
company					O
XYZ					Company
Private					Company
Limited					Company
works					O
in					O
the					O
health					Activity
sector					Activity
in					O
Europe					Location

# Clinical Word/Sentence Embeddings



# Clinical Word/Sentence Embeddings

Clinical Glove  
(200d)

PubMed + PMC

ICDO Glove  
(200d)

PubMed + ICD10  
UMLS + MIMIC III

Sent BERT

BioBert finetuned on  
NLI and MedNLI

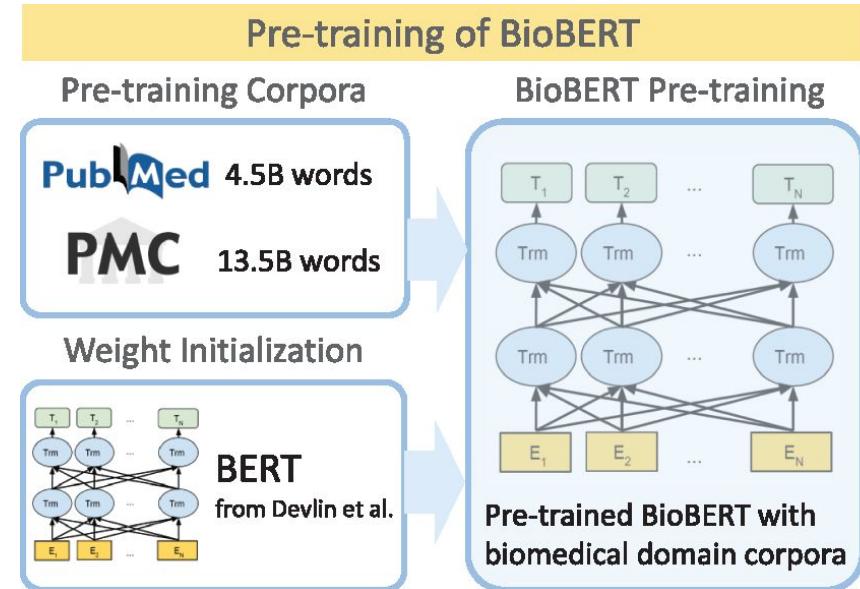
Bio/Clinical BERT

Fine tuned Pubmed + PMC + Discharge summaries



PubMed abstracts and PMC full-text articles

<https://www.nlm.nih.gov/bsd/difference.html>



# Part - II

- ❖ Assertion Status detection

# Assertion Status Detection

"Mother with a lung cancer, a patient is diagnosed as breast cancer in 1991 and then admitted to Mayo Clinic in Oct 2000, went under chemo for 6 months, discharged in April 2001 with a prescription of 2 mg metformin 3x per day. No sign of gynecological disorder but she suffers from acute cramps if she doesn't take her drug."

Chunk	Entity	Assertion
lung cancer	Oncological	Family
breast cancer	Oncological	Past
chemo	Treatment	Past
gynecological disorder	Disorder	Absent
acute cramps	Disorder	Conditional

```
clinical_assertion = AssertionDLModel\  
    .pretrained("assertion_dl", "en", "clinical/models")\  
    .setInputCols(["sentence", "ner_chunk", "embeddings"]) \  
    .setOutputCol("assertion")
```

Classify the assertions made on given medical concepts as being

- present,
- absent,
- possible,
- conditionally present under certain circumstances,
- hypothetically present at some future point, mentioned in the patient report but associated with someone else.

# Assertion Status Detection

- The deep neural network architecture for assertion status detection in Spark NLP is based on a Bi-LSTM framework, and is a modified version of the architecture proposed by Federico Fancellu, Adam Lopez and Bonnie Webber ([Neural Networks For Negation Scope Detection](#)).
- In the proposed implementation, input units depend on the target tokens (a named entity) and the neighboring words that are explicitly encoded as a sequence using word embeddings.
- Similar to paper mentioned above, it is observed that that 95% of the scope tokens (neighboring words) fall in a window of 9 tokens to the left and 15 to the right of the target tokens in the same dataset. Therefore, the same window size was implemented,
- following parameters were used: learning rate 0.0012, dropout 0.05, batch size 64 and a maximum sentence length 250.
- The model has been implemented within Spark NLP as an annotator called AssertionDLModel. After training 20 epoch and measuring accuracy on the official test set, this implementation exceeds the latest state-of-the-art accuracy benchmarks

Assertion Label	Spark NLP	Latest Best
Absent	0.944	0.937
Someone-else	0.904	0.869
Conditional	0.441	0.422
Hypothetical	0.862	0.890
Possible	0.680	0.630
Present	0.953	0.957
micro F1	0.939	0.934

Mother with a lung cancer,

# AssertionDLModel

	<b>model_name</b>	<b>Predicted Entities</b>
1	assertion_dl	Present, Absent, Possible, Planned, Someoneelse, Past, Family, None, Hypothetical
2	assertion_dl_biobert	absent, present, conditional, associated_with_someone_else, hypothetical, possible
3	assertion_dl_healthcare	absent, present, conditional, associated_with_someone_else, hypothetical, possible
4	assertion_dl_large	hypothetical, present, absent, possible, conditional, associated_with_someone_else
5	assertion_dl_radiology	Confirmed, Suspected, Negative
6	assertion_jsl	Present, Absent, Possible, Planned, Someoneelse, Past, Family, None, Hypothetical
7	assertion_jsl_large	present, absent, possible, planned, someoneelse, past
8	assertion_ml	Hypothetical, Present, Absent, Possible, Conditional, Associated_with_someone_else

# Spark NLP for Healthcare Data Scientists

July 20-21, 2022

**Veysel Kocaman**  
**Head of Data Science**  
[veysel@johnsnowlabs.com](mailto:veysel@johnsnowlabs.com)



# Part - III

- ❖ Entity Resolution (ICD1-, RxNorm, Snomed, etc.)

# Entity Resolution in Spark NLP for Healthcare

This is a 52-year-old AGE inmate with a 5.5 MEASUREMENTS cm UNITS diameter nonfunctioning mass SYMPTOM in his GENDER right DIRECTION adrenal BODYPART shown by CT of IMAGINGTEST abdomen BODYPART . During the umbilical hernia repair PROCEDURE , the harmonic scalpel MEDICAL\_DEVICE was utilised superiorly DIRECTION and laterally DIRECTION .

## Entity Resolution

ICD10CM, Snomed, RxNorm, CPT-4, ICD10CPS, RxCUI, ICDO

Term	Vocab	Code	Explanation (ground truth)
CT	CPT-4	76497	Unlisted computed tomography procedure
CT of abdomen	CPT-4	74150	Computed tomography, abdomen; without contrast material

## weighted Sentence Chunk Embeddings (after 3.2.0)

Term	Vocab	Code	Explanation (ground truth)
CT	CPT-4	74150	Computed tomography, abdomen; without contrast material

# Clinical Entity Resolution

## ICD10CM

- sbiobertresolve\_icd10cm\_augmented
- sbiobertresolve\_icd10pcs
- sbiobertresolve\_icd10cm\_augmented\_billable\_hcc
- sbiobertresolve\_icd10cm
- sbiobertresolve\_icd10cm\_slim\_normalized
- sbiobertresolve\_icd10cm\_slim\_billable\_hcc
- sbertrresolve\_icd10cm\_slim\_billable\_hcc\_med
- sbiobertresolve\_icd10cm\_generalised

## CPT

- sbiobertresolve\_cpt
- sbiobertresolve\_cpt\_procedures\_augmented
- sbiobertresolve\_cpt\_augmented
- sbiobertresolve\_cpt\_procedures\_measurements\_augmented

## Snomed

- sbiobertresolve\_snomed\_auxConcepts\_int
- sbiobertresolve\_snomed\_findings
- sbiobertresolve\_snomed\_findings\_int
- sbiobertresolve\_snomed\_auxConcepts
- sbertrsolve\_snomed\_bodyStructure\_med
- sbiobertresolve\_snomed\_bodyStructure
- sbiobertresolve\_snomed\_findings\_aux\_concepts
- sbertrsolve\_snomed\_conditions

## RxNorm

- sbiobertresolve\_rxnorm
- demo\_sbiobertresolve\_rxnorm
- sbiobertresolve\_rxnorm\_dispo
- sbiobertresolve\_rxnorm\_disposition
- sbertrsolve\_rxnorm\_disposition
- sbiobertresolve\_rxnorm\_ndc

## LOINC

- sbluebertresolve\_loinc
- sbiobertresolve\_loinc

and more ...

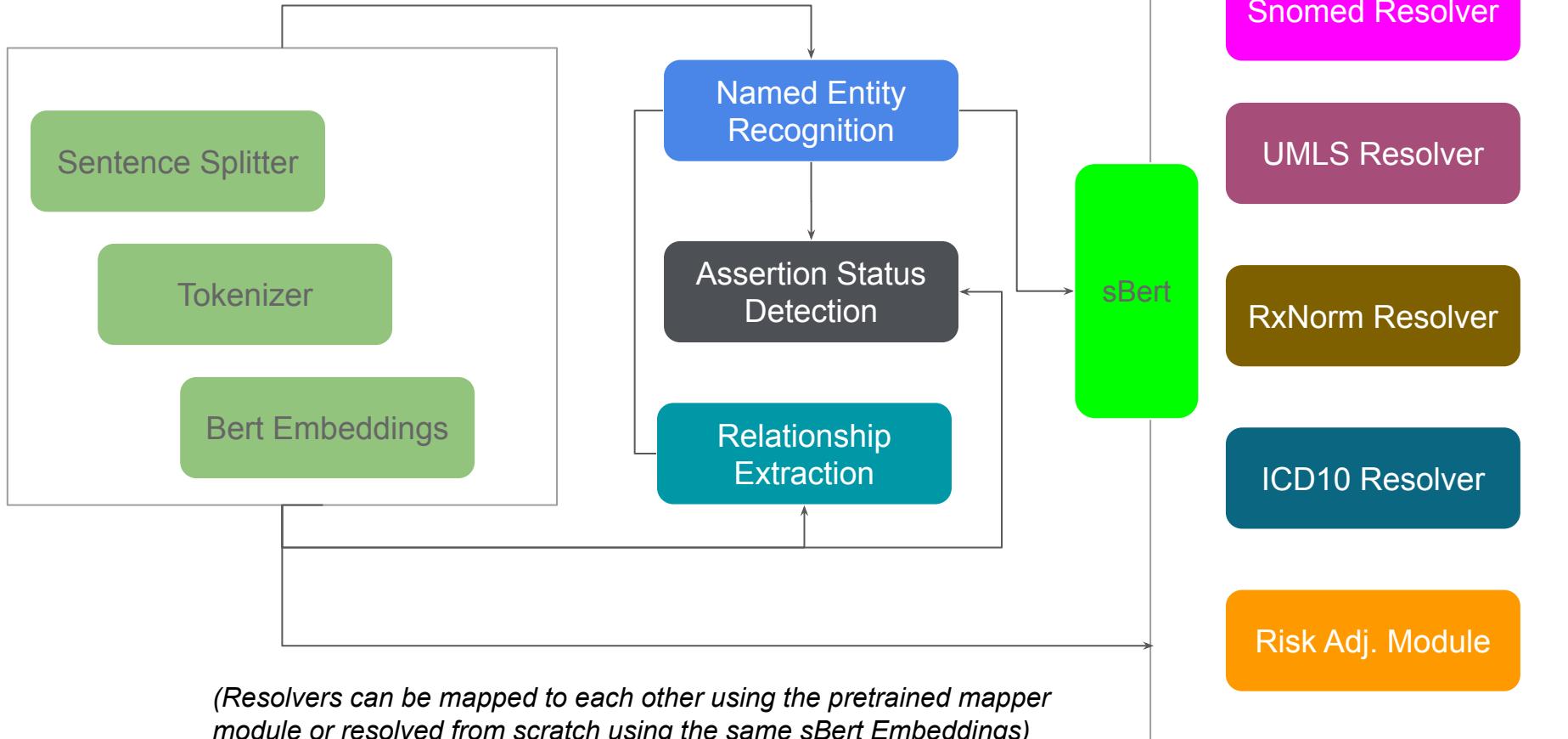
## UMLS

- sbiobertresolve\_umls\_findings
- sbiobertresolve\_umls\_major\_concepts
- sbiobertresolve\_umls\_disease\_syndrome
- sbiobertresolve\_umls\_clinical\_drugs

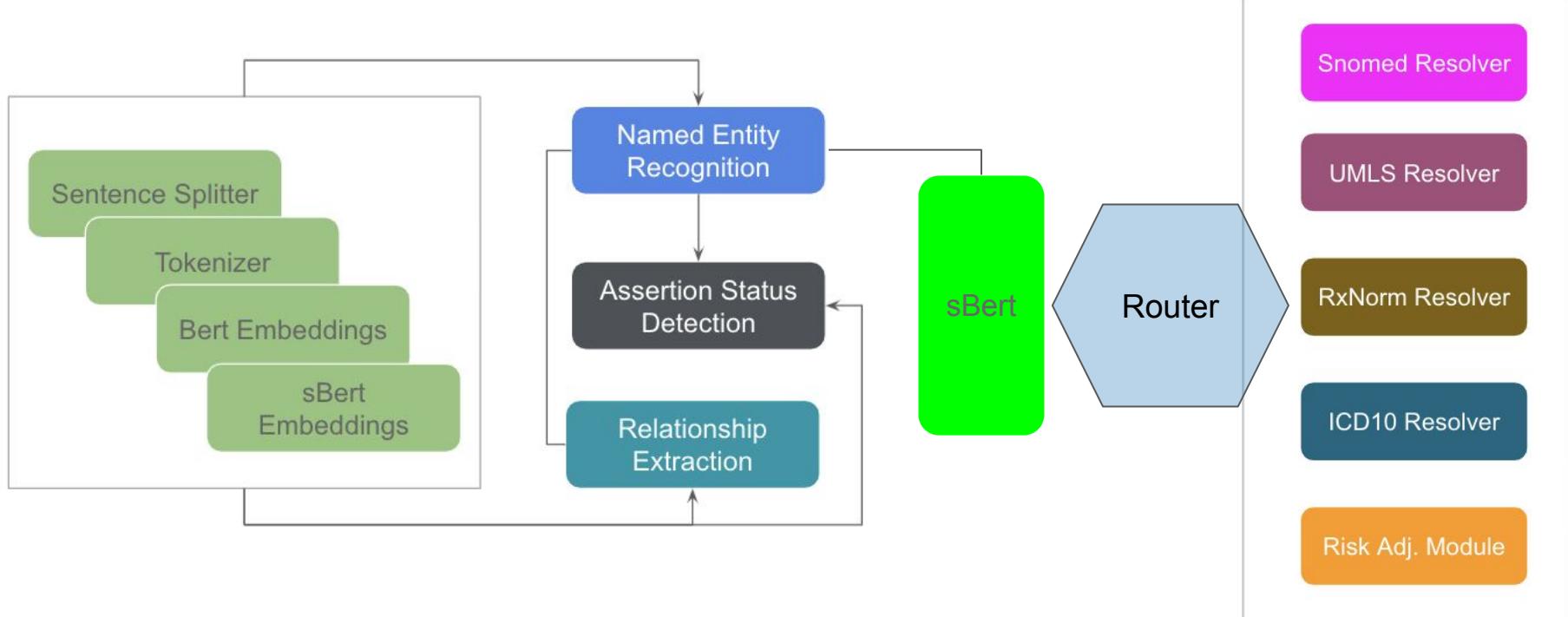
## mapping

- icd10cm\_snomed\_mapping : ICD10 Codes to Snomed Codes
- snomed\_icd10cm\_mapping : Snomed Codes to ICD Codes
- icd10cm\_umls\_mapping : ICD Codes to UMLS Codes
- snomed\_umls\_mapping : Snomed Codes to UMLS Codes
- rxnorm\_umls\_mapping : RxNorm Codes to UMLS Codes
- mesh\_umls\_mapping : MeSH Codes to UMLS Codes
- rxnorm\_mesh\_mapping : RxNorm Codes to MeSH Codes

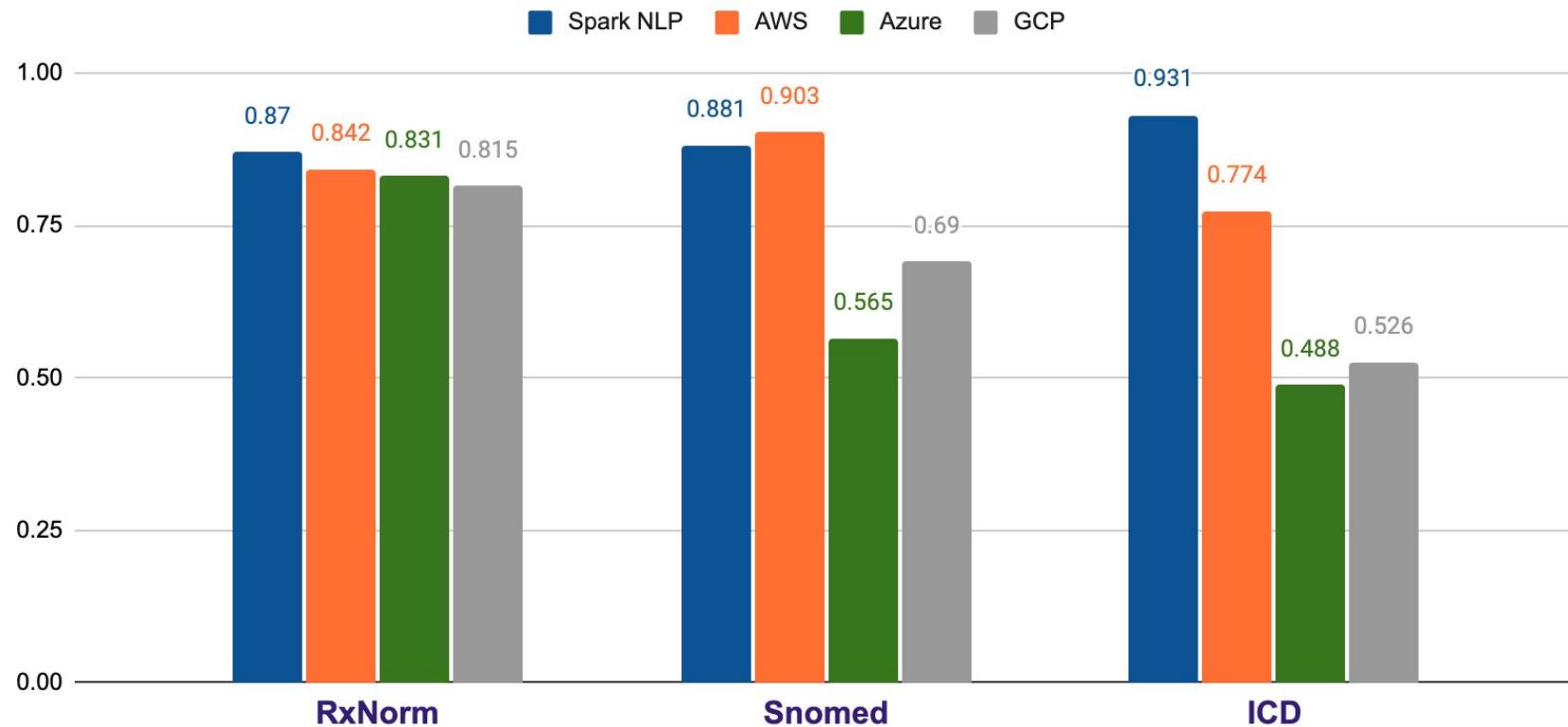
# Clinical Entity Resolution



# Clinical Entity Resolution



# Top - 5 Results



# Part - IV

- ❖ De-Identification and Obfuscation of PHI data

# Background & Motivation

Why de-identify?

- Organizations in possession of documents containing Protected Health Information (PHI) must follow privacy rules
- De-identification enables sharing of health data for medical research studies, policy assessments, other studies/assessments without violating patient privacy or requiring individual authorizations (HIPAA / GDPR restrictions around sharing no longer apply)

# Protected Health Information (PHI)

All info created or received by an entity that:

- Relates to past/present/future health or condition of an individual; the provision of health care; or the past/present/future payment for the provision of health care
- Identifies the individual; or with respect to which there is a reasonable basis to believe the information can be used to identify the individual

PHI includes many common identifiers: names, geographic locations, dates, phone numbers, email, IDs, SSN, medical record numbers, etc

# Spark NLP in Action

Spark NLP for Healthcare → De-Identification



**Deidentify structured  
data**

[Live Demo](#)[Colab Netbook](#)

**Deidentify free text  
documents**

[Live Demo](#)[Colab Netbook](#)

**Deidentify DICOM  
documents**

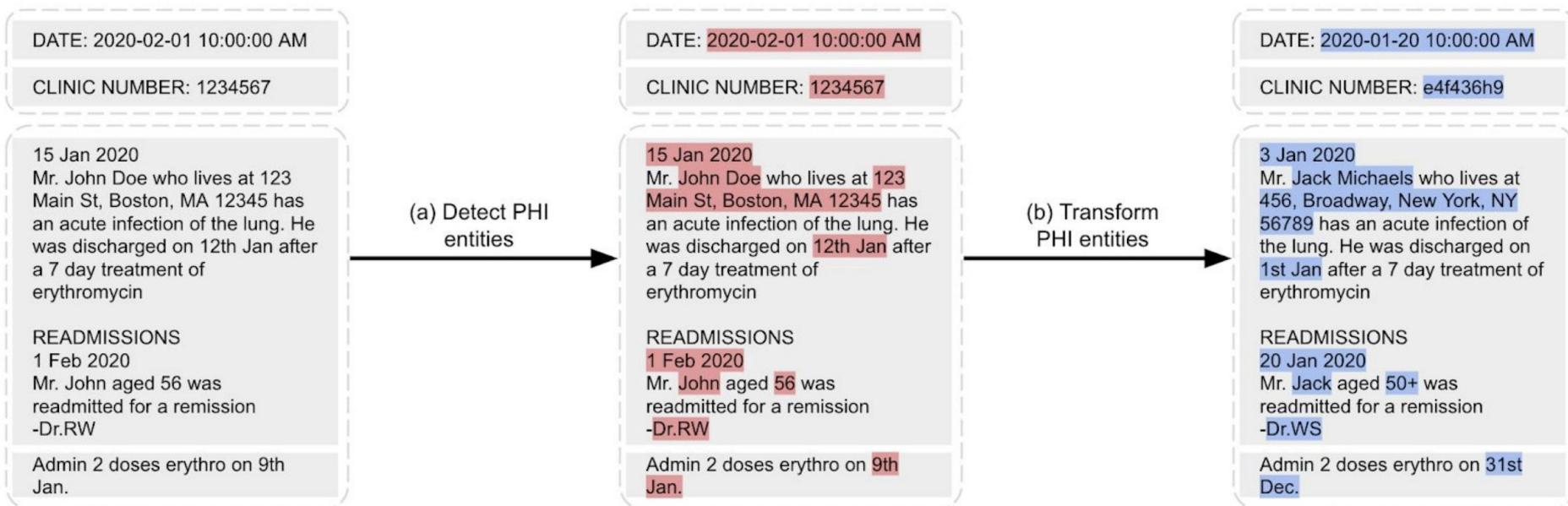
[Live Demo](#)[Colab Netbook](#)

**De-identify PDF  
documents - HIPAA  
Compliance**

[Live Demo](#)[Colab Netbook](#)

# De-Identification

\* Identifies potential pieces of content with personal information about patients and remove them by replacing with semantic tags.

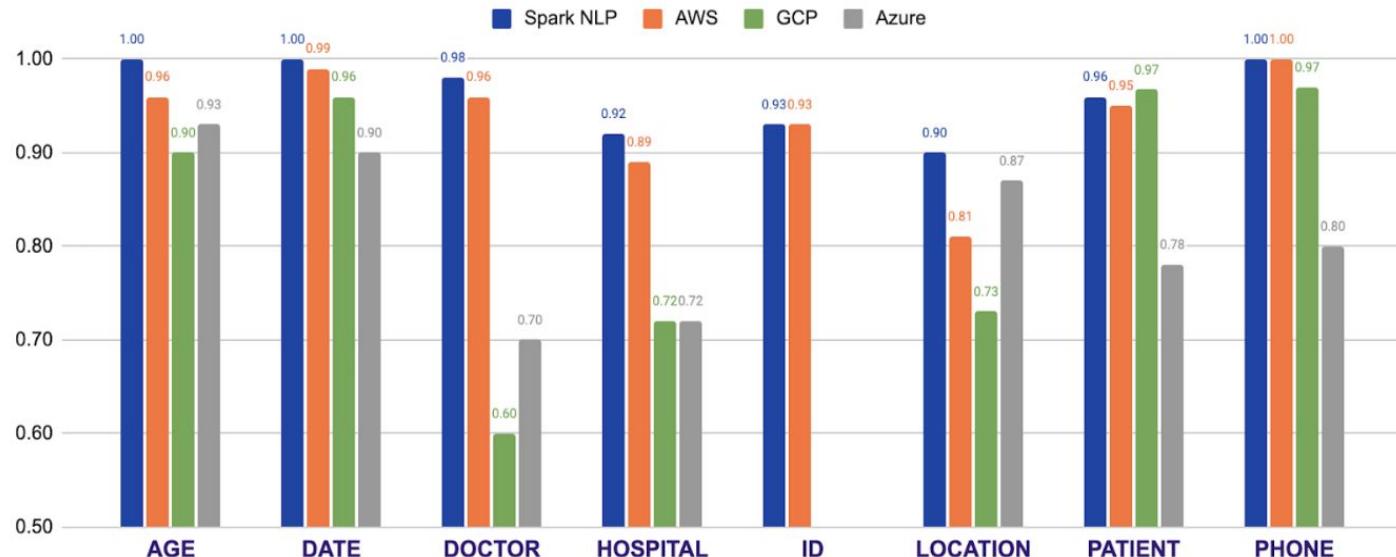


Veysel Kocaman

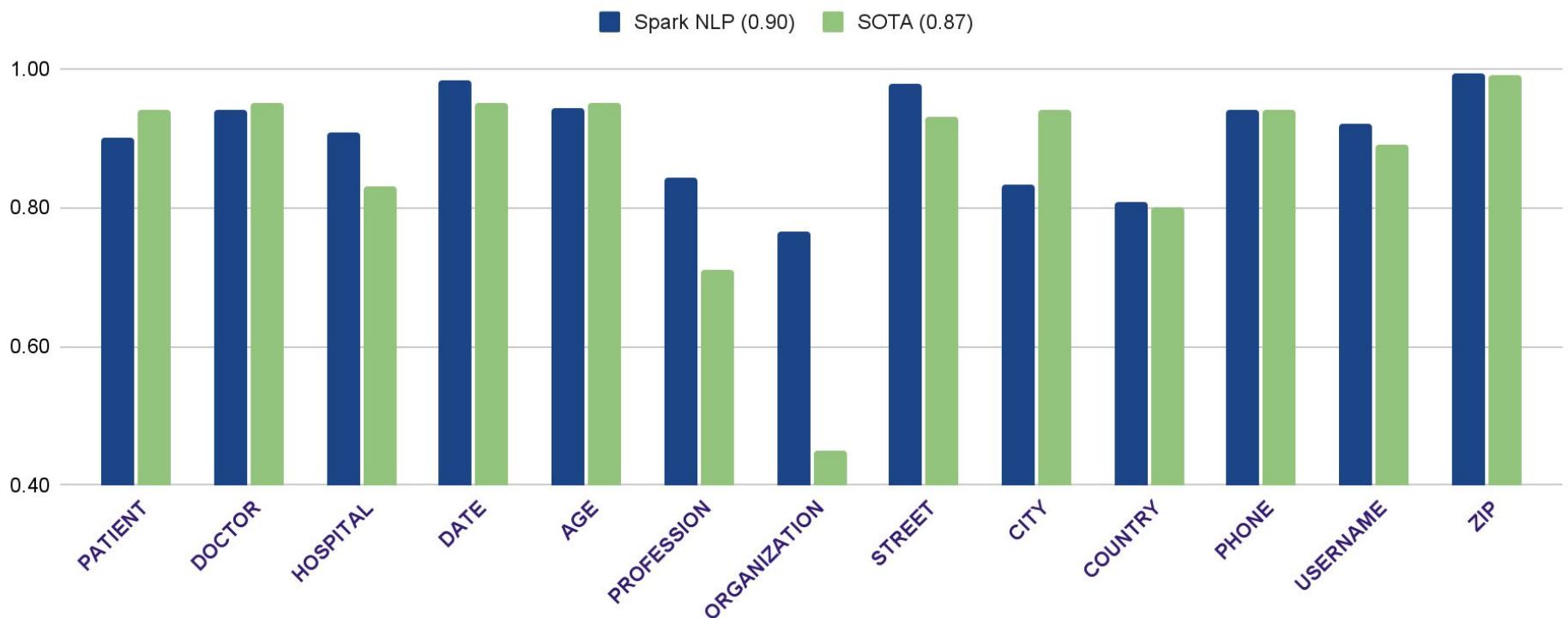
May 2 · 17 min read · [Listen](#)



# Comparison of Key Medical NLP Benchmarks — Spark NLP vs AWS, Google Cloud and Azure



# Deidentification Benchmarks





## How Providence Health De-Identified 700 Million Patient Notes with Spark NLP

Accuracy:

**99.19**

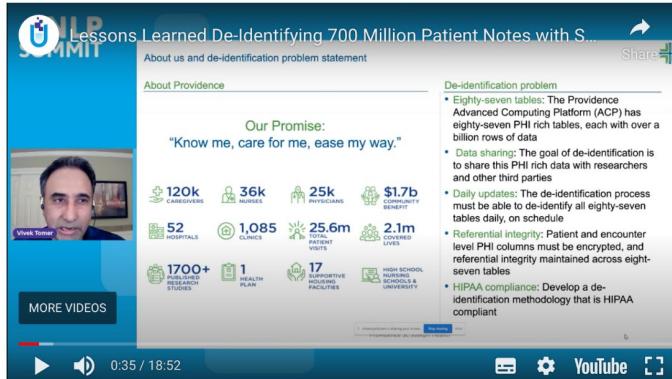
correctly de-identified sentences

Performance:

**2.46 hours**

to de-identify 500K patient notes.

[See how we did it](#)



**De-identify**



**Software**

**We support:**

- Tabular (headers, values)
- Text (NER, text matching)
- PDF: Text or Scanned
- Images (OCR & metadata)
- DICOM (OCR & metadata)

**So you can:**

- Replace (or delete a field)
- Mask (hash identifiers or shift dates)
- Obfuscate (name, locations, organizations)
- Generalize (disease codes, dates, addresses)

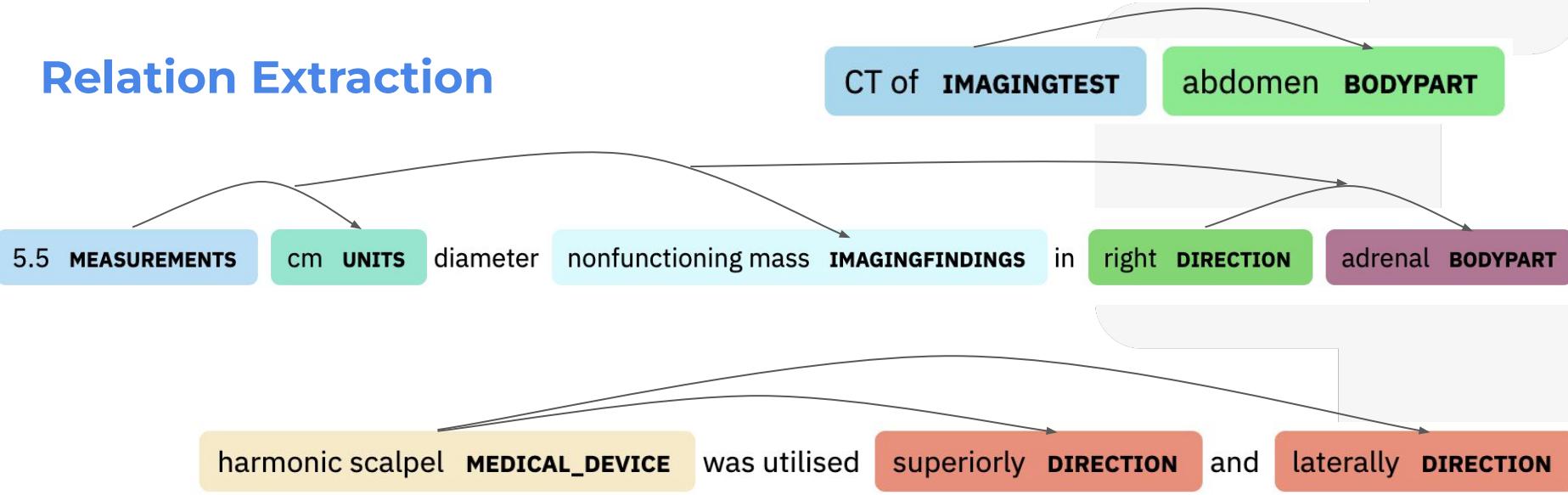
# Part - V

- ❖ Relation Extraction

# Clinical Relation Extraction

"This is a 52-year-old inmate with a 5.5 cm diameter nonfunctioning mass in his right adrenal shown by CT of abdomen. During the umbilical hernia repair, the harmonic scalpel was utilised superiorly and laterally."

## Relation Extraction



# Clinical Relation Extraction

## model\_name

0	re_ade_biobert
1	re_ade_clinical
2	re_bodypart_directions
3	re_bodypart_problem
4	re_bodypart_proceduretest
5	re_chemprot_clinical
6	re_clinical
7	re_date_clinical
8	re_drug_drug_interaction_clinical
9	re_human_phenotype_gene_clinical
10	re_temporal_events_clinical
11	re_temporal_events_enriched_clinical
12	re_test_problem_finding
13	re_test_result_date

14	redl_ade_biobert
15	redl_bodypart_direction_biobert
16	redl_bodypart_problem_biobert
17	redl_bodypart_procedure_test_biobert
18	redl_chemprot_biobert
19	redl_clinical_biobert
20	redl_date_clinical_biobert
21	redl_drug_drug_interaction_biobert
22	redl_human_phenotype_gene_biobert
23	redl_temporal_events_biobert

Relation	Recall	Precision	F1	SOTA
DRUG-ADE	0.66	1.00	<b>0.80</b>	0.76
DRUG-DOSAGE	0.89	1.00	<b>0.94</b>	0.91
DRUG-DURATION	0.75	1.00	<b>0.85</b>	0.92
DRUG-FORM	0.88	1.00	<b>0.94</b>	0.95*
DRUG-FREQUENCY	0.79	1.00	<b>0.88</b>	0.90
DRUG-REASON	0.60	1.00	<b>0.75</b>	0.70
DRUG-ROUTE	0.79	1.00	<b>0.88</b>	0.95*
DRUG-STRENGTH	0.95	1.00	<b>0.98</b>	0.97

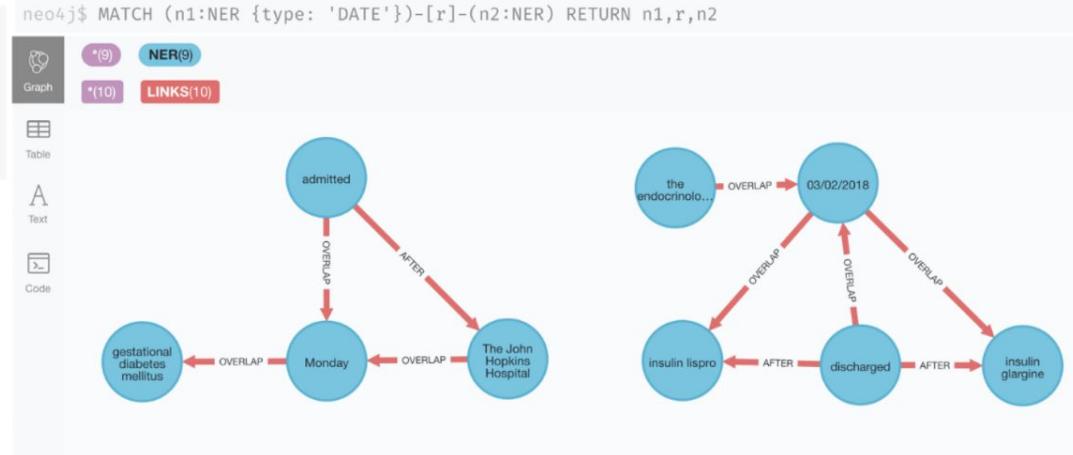
Relation	Recall	Precision	F1
OVERLAP	0.81	0.73	<b>0.77</b>
BEFORE	0.85	0.88	<b>0.86</b>
AFTER	0.38	0.46	<b>0.43</b>

# Clinical Relation Extraction

*She is admitted to The John Hopkins Hospital on Monday with a history of gestational diabetes mellitus diagnosed. She was seen by the endocrinology service and she was discharged on 03/02/2018 on 40 units of insulin glargin and 12 units of insulin lispro.*

```
1 query = """
2 | MATCH (n1:NER {type: 'DATE'})-[r]-(n2:NER)
3 | RETURN n1.name AS date, r.relation AS relation, n2.name AS event
4 """
5
6 df = pd.DataFrame([dict(_) for _ in conn.query(query)])
7 df
```

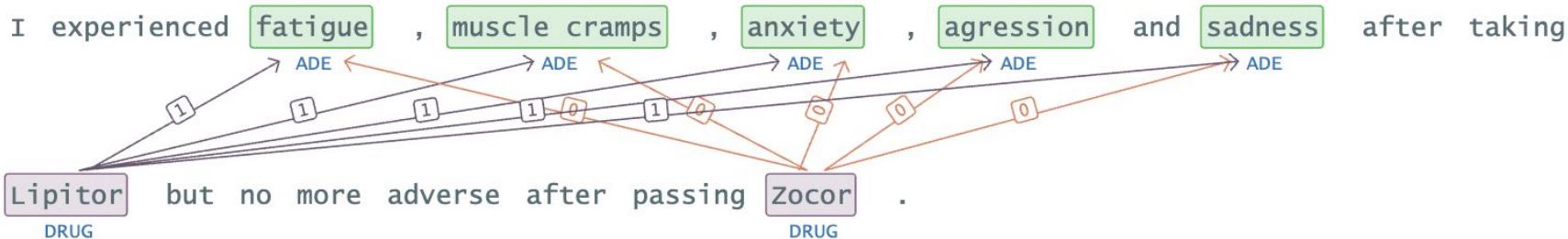
	date	relation	event
0	Monday	OVERLAP	gestational diabetes mellitus
1	Monday	OVERLAP	The John Hopkins Hospital
2	Monday	OVERLAP	admitted
3	03/02/2018	OVERLAP	insulin lispro
4	03/02/2018	OVERLAP	insulin glargin
5	03/02/2018	OVERLAP	discharged
6	03/02/2018	OVERLAP	the endocrinology service



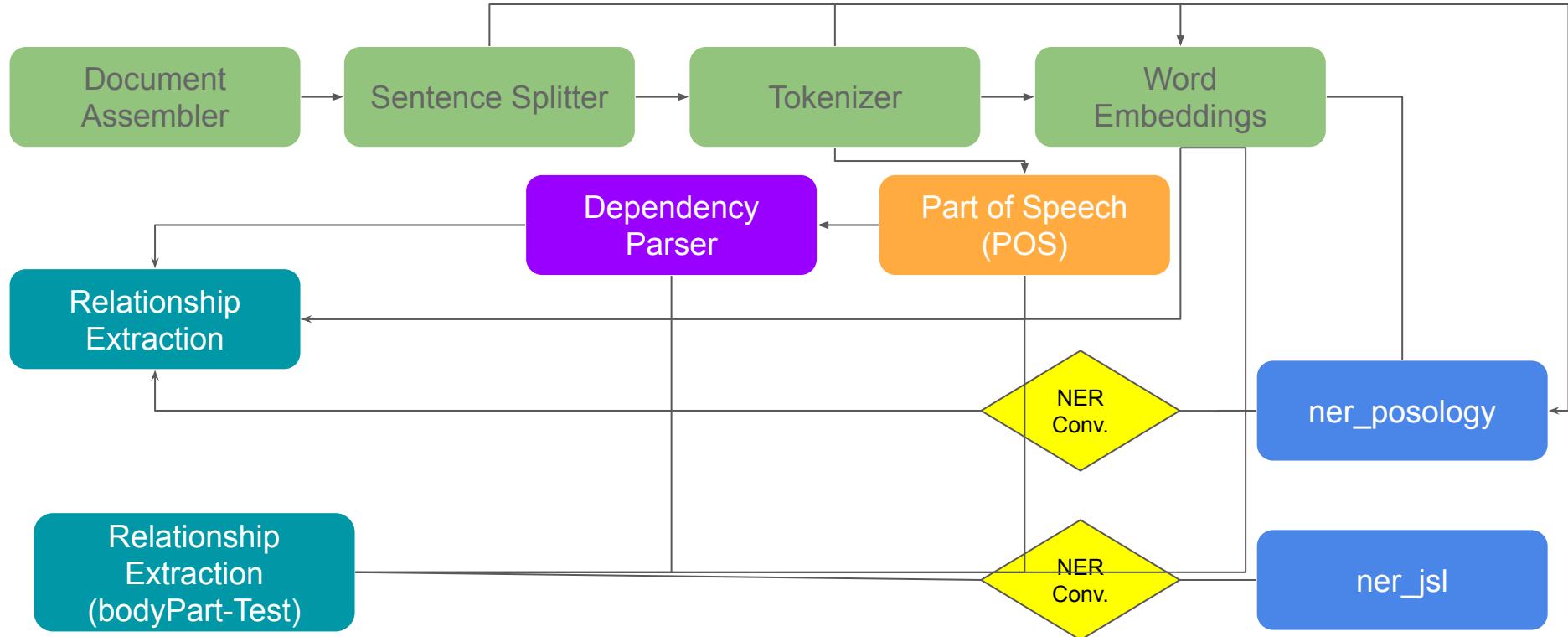
# Clinical Relation Extraction

	relation	entity1	entity1_begin	entity1_end	chunk1	entity2	entity2_begin	entity2_end	chunk2	confidence
0	1	ADE	14	20	fatigue	DRUG	82	88	Lipitor	0.9996617
1	0	ADE	14	20	fatigue	DRUG	124	128	Zocor	0.9952187
2	1	ADE	23	35	muscle cramps	DRUG	82	88	Lipitor	0.9999827
3	0	ADE	23	35	muscle cramps	DRUG	124	128	Zocor	0.91462934
4	1	ADE	38	44	anxiety	DRUG	82	88	Lipitor	0.7636133
5	0	ADE	38	44	anxiety	DRUG	124	128	Zocor	0.9999691
6	1	ADE	47	55	agression	DRUG	82	88	Lipitor	0.99999833
7	0	ADE	47	55	agression	DRUG	124	128	Zocor	0.99781835
8	1	ADE	61	67	sadness	DRUG	82	88	Lipitor	1.0
9	0	ADE	61	67	sadness	DRUG	124	128	Zocor	0.9999572

I experienced fatigue, muscle cramps, anxiety, agression and sadness after taking Lipitor but no more adverse after passing Zocor.



# Clinical Relation Extraction

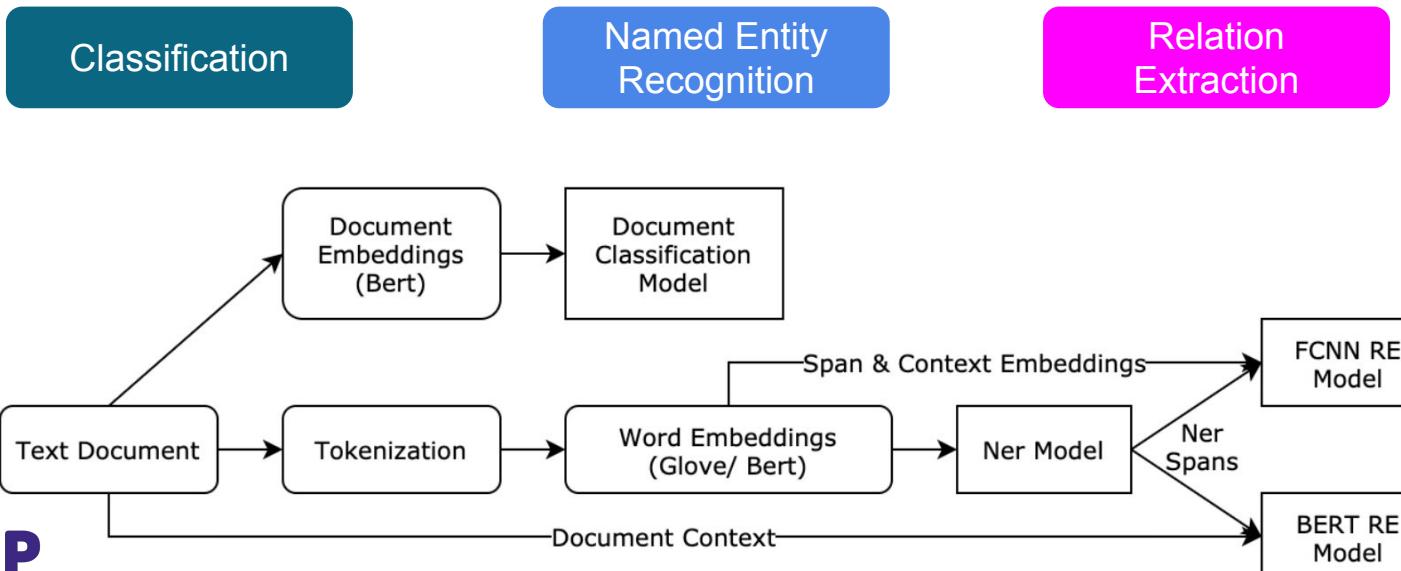


# Part - VI

## ❖ Adverse Drug Reactions (ADR)

# Adverse Drug Reactions (ADR)

Document	Class	ADE Entity	Drug Entity	relation
I feel a bit drowsy & have a little blurred vision after taking insulin.	ADE	drowsy blurred vision	insulin insulin	Positive Positive
@yho fluvastatin gave me cramps, but lipitor suits me!	ADE	cramps cramps	fluvastatin lipitor	Positive Negative
I just took advil and haven't had any gastric problems so far.	NEG	-	-	-



# Adverse Drug Reactions (ADR) Benchmark



Dataset	GLoVe Embeddings						BERT Embeddings						SOTA	
	Precision		Recall		F1		Precision		Recall		F1		F1	
	strict	relax	strict	relax	strict	relax	strict	relax	strict	relax	strict	relax		
ADE	88.32	93.77	89.26	94.80	88.78	94.27	90.0	94.47	93.56	98.22	<b>91.75</b>	96.31	<b>91.3</b>	
	87.81	93.59	88.81	94.66	88.30	94.12	89.6	94.37	93.18	98.13	91.36	96.21		
CADEC	78.14	89.04	77.14	88.01	77.62	88.50	78.53	88.63	79.03	89.32	<b>78.76</b>	88.95	<b>71.9</b>	
	71.87	86.36	71.67	86.13	71.75	86.23	72.38	86.14	73.64	87.66	72.99	86.88		
SMM4H	81.43	90.33	72.17	78.51	76.01	83.41	78.5	86.76	75.23	82.42	<b>76.73</b>	84.41	<b>67.81</b>	
	83.66	91.34	71.31	77.86	76.99	84.06	79.13	87.09	74.33	81.81	76.65	84.36		

Table 2: NER metrics on benchmark datasets. For each dataset, macro and micro averaged scores are displayed on first and second row respectively. SOTA metrics for ADE, CADEC, and SMM4H are obtained from (Yan et al. 2021), (Stanovsky, Gruhl, and Mendes 2017), and (Ge et al. 2020) respectively, and are macro-averaged.

## Named Entity Recognition (NER)

# Adverse Drug Reactions (ADR) Benchmark



Dataset	GLoVe (Avg.) Embeddings			BERT (Avg.) Embeddings			BERT Sentence Embeddings			SOTA
	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1	F1
ADE	75.96	79.53	76.86	76.91	84.96	79.37	87.41	84.72	<b>85.96</b>	<b>87.0</b>
	86.84	81.22	83.43	88.13	84.38	85.38	90.97	91.20	91.03	
CADEC	85.29	84.24	84.71	86.50	86.11	86.30	87.13	86.32	<b>86.69</b>	<b>81.5</b>
	85.99	86.10	86.0	87.38	87.43	87.40	87.78	87.86	87.79	

Table 1: Classification Metrics on benchmark datasets. For each dataset, Macro and Micro averaged scores are displayed on first and second row respectively. SOTA metrics for ADE and CADEC datasets are obtained from (Huynh et al. 2016) and (Alimova and Tutubalina 2019) respectively.

Dataset	Base (FCNN) RE			BERT RE			SOTA
	Precision	Recall	F1	Precision	Recall	F1	F1
ADE Corpus	69.11	86.22	<b>74.70</b>	81.31	79.03	<b>80.10</b>	<b>83.74</b>
ADE Enriched with n2c2	89.01	89.44	<b>89.22</b>	89.19	90.93	<b>90.02</b>	

Table 3: Relation Extraction performance on the ADE benchmark dataset. The test set was kept standard for a fair comparison, and all scores are macro-averaged due to high class imbalance. SOTA metrics for RE on ADE corpus as reported by (Crone 2020)

# Part - VII

## ❖ Key Chunk Phrase Extractor

Chunk KeyPhrase Extraction uses Bert Sentence Embeddings to determine the most relevant key phrases describing a text. The input to the model consists of chunk annotations and sentence or document annotation. The model compares the chunks against the corresponding sentences/documents and selects the chunks which are most representative of the broader text context (i.e. the document or the sentence they belong to). The key phrases candidates (i.e. the input chunks) can be generated in various ways, e.g. by NGramGenerator, TextMatcher or NerConverter. The model operates either at sentence (selecting the most descriptive chunks from the sentence they belong to) or at document level. In the latter case, the key phrases are selected to represent all the input document annotations.

# Chunk Key Phrase Extractor (KCPE)

key_phrase	source	DocumentSimilarity	MMRScore	sentence
type two diabetes mellitus	NER	0.7639750686118073	0.4583850593816694	0
subsequent type diabetes	ngrams	0.7503709443591438	0.08298243928224425	0
HTG-induced pancreatitis years	ngrams	0.6817062970203589	0.11246275270031031	0
hepatitis obesity	ngrams	0.6666053470245074	0.1177052008980295	0
mellitus diagnosed years	ngrams	0.6389213391545323	0.08129479185432026	0
history gestational diabetes	ngrams	0.6219876368539883	0.0950104202982544	0
vomiting	ngrams	0.5824238088130589	0.14864183399720493	0
admitted starvation ketosis	ngrams	0.5789875069392564	0.12008073486190007	0
five-day amoxicillin respiratory	ngrams	0.5330653868257814	0.09428153526023508	0
28-year-old female history	ngrams	0.38613601247069695	0.12987678861407687	0

## YAKE

keyword	score
years prior presentation	0.006335399690627251
prior presentation	0.011644010991495998
prior presentation subsequent	0.020272229518351368
weeks prior presentation	0.020272229518351368
respiratory tract infection	0.02568455658449274
anion gap	0.025965846371439553
physical examination presentation	0.02840600503736659
obtained hours presentation	0.028532992974589392
examination presentation significant	0.028532992974589392
prior	0.029673513395379065
years prior	0.03008818777992058
anion gap elevated	0.031568192739369824

## CKPE

key_phrase_candidate	DocumentSimilarity
pancreatitis years prior	0.6491587146812722
diagnosed years prior	0.38594469396979897
respiratory tract infection	0.34452766290310755
patient treated insulin	0.3413457416284759
serum	0.3371024001999838
presentation revealed glucose	0.31458360368143906
examination presentation significant	0.29099950377907047
prior analysis due	0.22501711661945623
prior	0.21634008371261446
physical examination presentation	0.19165189487112474

key_phrase_candidate	source
28-year-old female history	ngram
28-year-old	NER
female history gestational	ngram
female	NER
history gestational diabetes	ngram
gestational diabetes mellitus	NER
gestational diabetes mellitus	ngram
diabetes mellitus diagnosed	ngram
mellitus diagnosed years	ngram
diagnosed years prior	ngram
eight years prior	NER
years prior presentation	ngram
prior presentation subsequent	ngram
presentation subsequent type	ngram
subsequent type diabetes	ngram
type diabetes mellitus	ngram
type two diabetes mellitus	NER
diabetes mellitus (	ngram
mellitus ( T2DM	ngram
( T2DM ),	ngram
T2DM ), prior	ngram
T2DM	NER
), prior episode	ngram
prior episode HTG-induced	ngram
episode HTG-induced pancreatitis	ngram
HTG-induced pancreatitis years	ngram
HTG-induced pancreatitis	NER
pancreatitis years prior	ngram
three years prior	NER
years prior presentation	ngram
prior presentation ,	ngram
presentation , acute	ngram
, acute hepatitis	ngram



# Thank you !

**Veysel Kocaman**  
Head of Data Science  
John Snow Labs



# Doc2Chunk / Chunk2Doc

- ✓ Transform the data from one AnnotatorType to another.
- ✓ Doc2Chunk
  - Converts DOCUMENT type annotations into CHUNK type with the contents of a chunkCol.
- ✓ Chunk2Doc
  - Converts a CHUNK type column back into DOCUMENT.
  - Useful when trying to re-tokenize or do further analysis on a CHUNK result.

```
chunkAssembler = Doc2Chunk() \  
  .setInputCols("document") \  
  .setChunkCol("target") \  
  .setOutputCol("chunk")
```

```
chunkToDoc = Chunk2Doc() \  
  .setInputCols("chunk") \  
  .setOutputCol("chunkConverted")
```

# Spark NLP Resources

Spark NLP Official page

Spark NLP Workshop Repo

JSL Youtube channel

JSL Blogs

Introduction to Spark NLP: Foundations and Basic Components (Part-I)

Introduction to: Spark NLP: Installation and Getting Started (Part-II)

Named Entity Recognition with Bert in Spark NLP

Text Classification in Spark NLP with Bert and Universal Sentence Encoders

Spark NLP 101 : Document Assembler

Spark NLP 101: LightPipeline

<https://www.oreilly.com/radar/one-simple-chart-who-is-interested-in-spark-nlp/>

<https://blog.dominodatalab.com/comparing-the-functionality-of-open-source-natural-language-processing-libraries/>

<https://databricks.com/blog/2017/10/19/introducing-natural-language-processing-library-apache-spark.html>

<https://databricks.com/fr/session/apache-spark-nlp-extending-spark-ml-to-deliver-fast-scalable-unified-natural-language-processing>

<https://medium.com/@saif1988/spark-nlp-walkthrough-powered-by-tensorflow-9965538663fd>

<https://www.kdnuggets.com/2019/06/spark-nlp-getting-started-with-worlds-most-widely-used-nlp-library-enterprise.html>

<https://www.forbes.com/sites/forbestechcouncil/2019/09/17/why-spark-nlp-is-the-most-widely-used-nlp-library-enterprise/>

<https://medium.com/hackernoon/mueller-report-for-nerds-spark-meets-nlp-with-tensorflow-and-bert-part-1-32490a8f8f12>

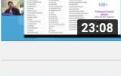
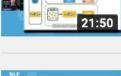
<https://www.analyticsindiamag.com/5-reasons-why-spark-nlp-is-the-most-widely-used-library-enterprise/>

<https://www.oreilly.com/ideas/comparing-production-grade-nlp-libraries-training-spark-nlp-and-spacy-pipelines>

<https://www.oreilly.com/ideas/comparing-production-grade-nlp-libraries-accuracy-performance-and-scalability>

<https://www.infoworld.com/article/3031690/analytics/why-you-should-use-spark-for-machine-learning.html>

# Healthcare NLP Summit '22

1	 End-to-End No-Code Development of NER model for Text with Annotation Lab John Snow Labs  4:42	Using Spark NLP in R: A Drug Standardization Case Study John Snow Labs  15:00	11	Natural Language Technologies: Current Status and Future Evolution John Snow Labs  19:54
2	 How to Build a Foundation of AI-based Healthcare Systems Through Language Models? John Snow Labs  34:12	How to sign up for a free trial and how to use your license on COLAB John Snow Labs  1:09	12	The Quest for Proactive and Reactive Healthcare John Snow Labs  15:31
3	 Automated Patient Risk Adjustment and Medicare HCC Coding from Clinical Notes John Snow Labs  28:19	Few-Shot Text Classification in the Real-World John Snow Labs  25:16	13	The Unified NLP Platform John Snow Labs  26:44
4	 End-to-End No-Code Development of Visual NER Models for PDFs and Images John Snow Labs  6:31	Medical NLP: Domain Expertise and Data Quality are Vital For Success John Snow Labs  13:50	14	Industry Survey Analysis: AI in Healthcare 2022 John Snow Labs  19:59
5	 A Hierarchical Approach for Automated ICD-10 Coding Using Phrase-level Attention John Snow Labs  29:04	Automatic mining of adverse drug reactions from social media posts and unstructured chats John Snow Labs  19:24	15	Using NLP at Scale to Process Patient Charts for Identifying Patient Encounters John Snow Labs  23:54
6	 End-to-End No-Code Development of AI Models for Text and Images John Snow Labs  14:41	Data Centric AI for Healthcare John Snow Labs  26:23	16	Transfer Learning From Existing Diseases Via Hierarchical Multi-Modal BERT Models to Predict COVID19 John Snow Labs  24:19
7	 Opportunities and Challenges of Applying Advances in NLP to Healthcare John Snow Labs  20:38	Radiology Report Summarization John Snow Labs  36:15	17	Supporting Mental Healthcare Delivery Using NLP John Snow Labs  25:33
8	 Current State-of-the-Art Accuracy for Key Medical Natural Language Processing Benchmarks John Snow Labs  23:08	Entropy and Sentiment: the Anna Karenina Principle in Patient Experience Data John Snow Labs  31:06	18	Using Spark NLP to De-Identify Doctor Notes in the German Language John Snow Labs  28:50
9	 Machine Reading for Precision Medicine John Snow Labs  21:50	1 Line of Code to Use 600+ State-of-the-Art Clinical & Biomedical NLP Models John Snow Labs  31:31	19	Accelerating Use of the Digital Medical Record with Optimized Structured & Unstructured Data John Snow Labs  30:20
10	 Collaborative Healthcare NLP: Customisable NLP platforms for health and related research John Snow Labs  18:10	Cleanlab: Making AI Work with Messy, Real-World Healthcare and NLP Data John Snow Labs  23:40	20	Is it Enough to Simply Apply Language Model for Optimal Text Classification? John Snow Labs  25:57