Spark NLP for Healthcare: Modular Approach to Solve Problems at Scale in Healthcare NLP

April 20, 2021

ODSC EAST
BOSTON I APRIL 19-21

Colab notebook : https://bit.ly/spark_nlp_colab
License keys : https://bit.ly/spark_nlp_license
Slides : https://bit.ly/spark_nlp_slides

Veysel Kocaman
Lead Data Scientist
John Snow Labs

The Challenges we face today;

Healthcare AI practitioners usually need to make unacceptable trade-offs between

- delivering state-of-the-art accuracy,
- generalizing over unseen data points,
- preventing the sharing of personal data or intellectual property.

NLP in Healthcare

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:

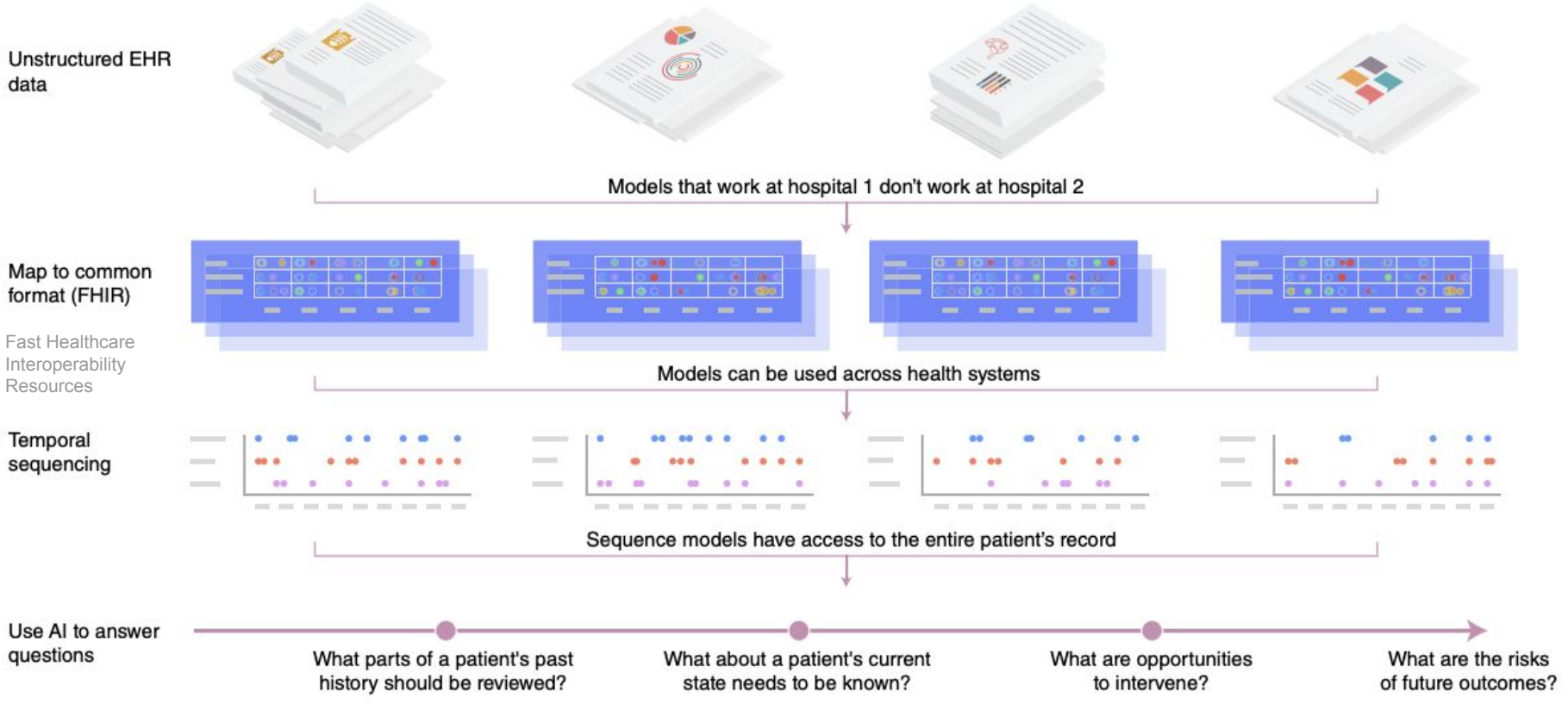
- Nuanced
- Fuzzy
- Contextual
- Medium specific
- Domain specific

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 yeatreday 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

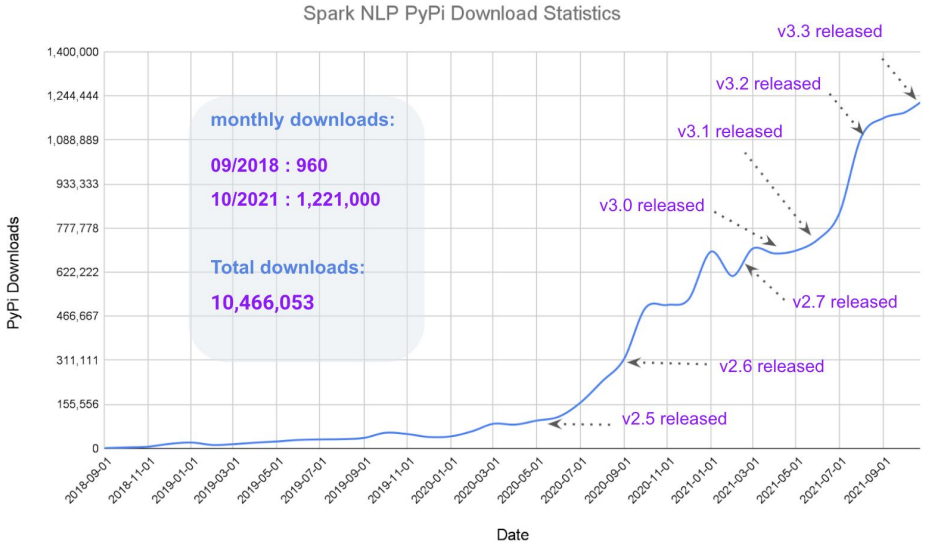


"Systems used to generate health data are designed for operations, not to organize data effectively for research or analytics."

Introducing Spark NLP

Daily ~ 50K
Monthly ~ 4.2M

PyPI link	https://pypi.org/project/spark-nlp
Total downloads	24,024,865
Total downloads - 30 days	4,169,282
Total downloads - 7 days	377,714



- 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)
  - Delivering a mission-critical, enterprise grade NLP library (used by multiple Fortune 500)
  - Full-time development team (a new release every other week)

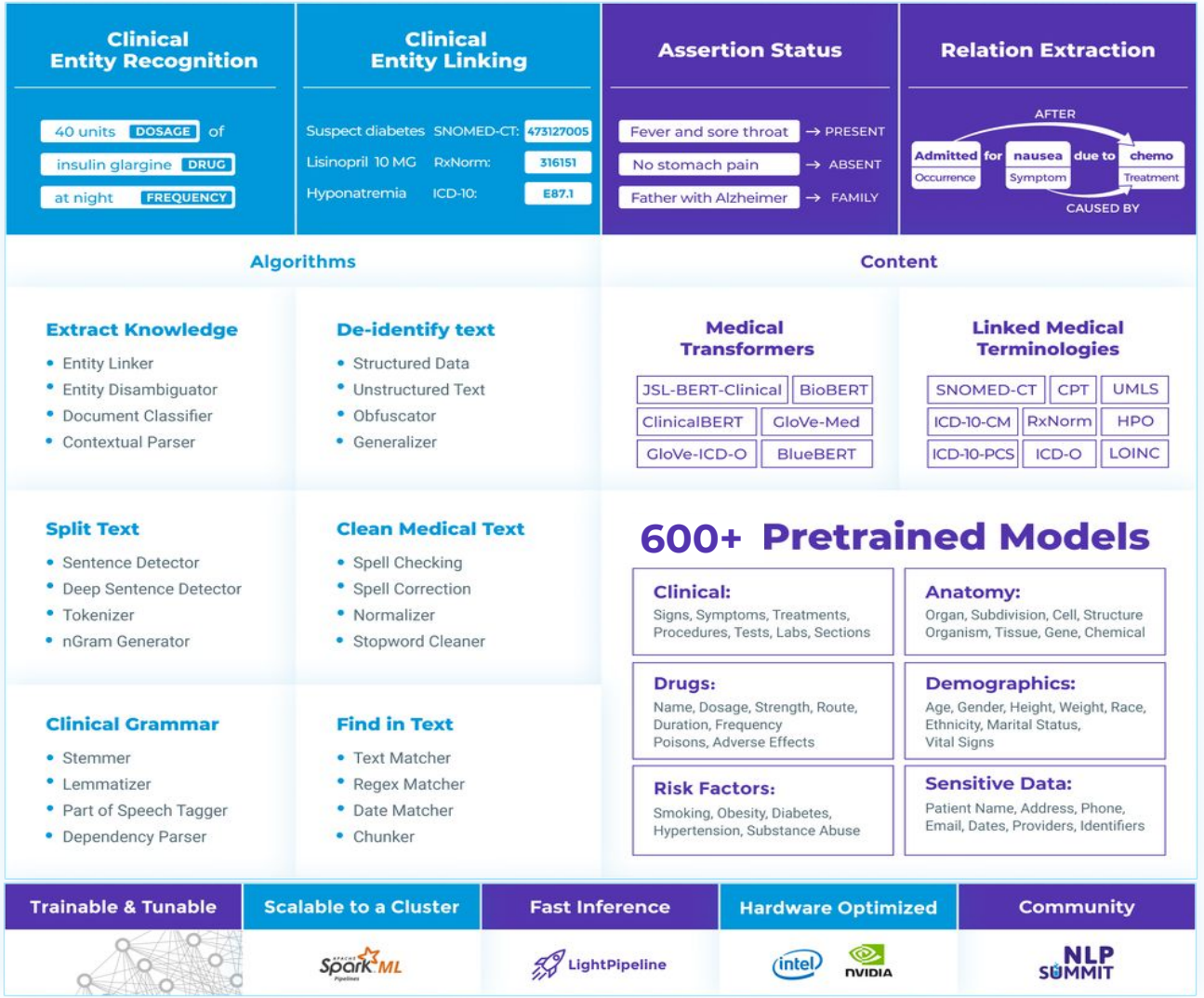
John Snow LABS

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.



Academic Activities & Benchmarks

John Snow LABS

nature COMMUNICATIONS

Stanford University

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

Khushbu Agarwal[1], Sutanay Choudhury*[1], Sindhu Tipirneni[2], Pritam Mukherjee[3], Colby Ham[1], Suzanne Tamang[4], Matthew Baker[5], Siyi Tang[5], Veysel Kocaman[7], Olivier Gevaert[3,4], Robert Rallo[1], and Chandan K Reddy[2]

[1]Pacific Northwest National Laboratory, Richland, 99354, USA
[2]Department of Computer Science, Virginia Tech, Arlington, 22203, USA
[3]Stanford Center for Biomedical Informatics Research, Department of Medicine, School of Medicine, Stanford University, Stanford, 94305, USA
[4]Department of Biomedical Data Science, Stanford University, Stanford, 94305, USA
[5]Department of Electrical Engineering, Stanford University, Stanford, 94305, USA
[6]Division of Immunology and Rheumatology, Department of Medicine, Stanford University, Stanford, 94305, USA
John Snow Labs, Delaware City, 19958, USA
sutanay.choudhury@pnnl.gov

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 as-

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-

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

Health Intelligence (W3PHIAI-22) at AAAI

ADR's has been estimated to cost $136 billion each year in the United States alone (van Der Hooft 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,

AMIA 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[1], Khushbu Agarwal[1], Colby Ham[1], Pritam Mukherjee[2], Siyi Tang[2], Sindhu Tipirneni[3], Chandan Reddy[3], Suzanne Tamang[2], Robert Rallo[1], Veysel Kocaman[4]
[1]Pacific Northwest National Laboratory; [2]Stanford University; [3]Virginia Tech.; [4]John Snow Labs

Introduction

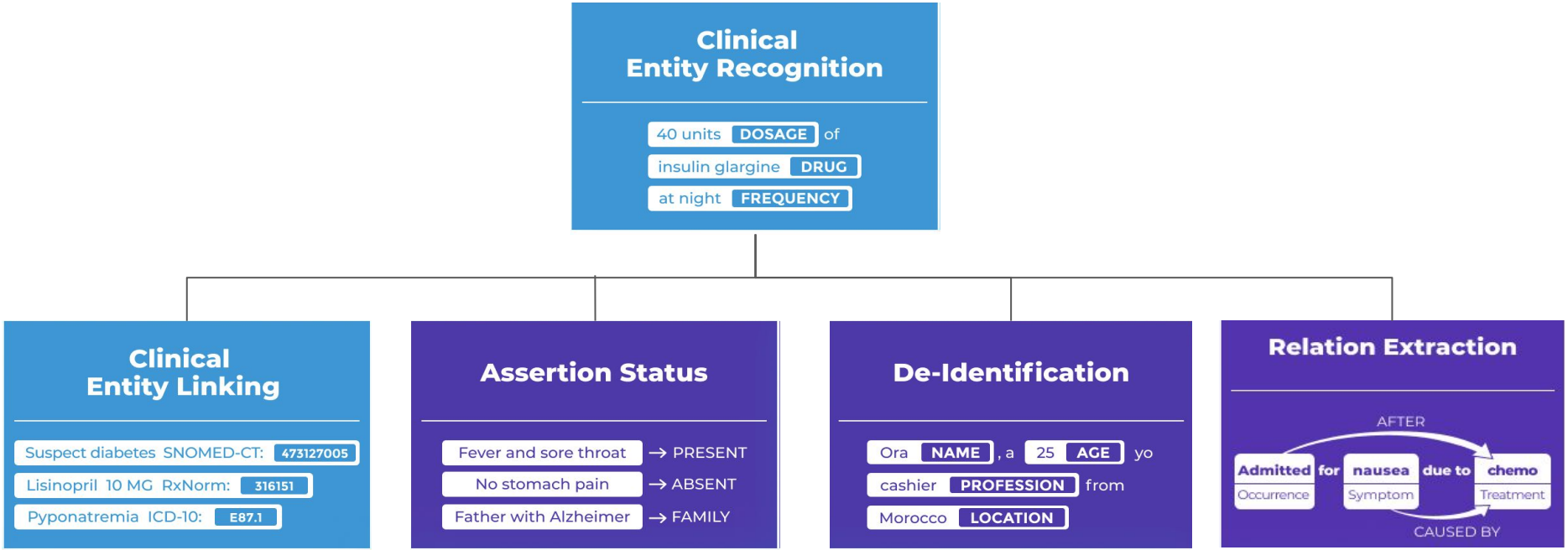
We aim to characterize the evolution in the effectiveness of treatment for different patient groups over the course of the COVID-19 pandemic. In contrast to most existing studies[1], we study the evolution of patient trajectories based on unique sets of frequent comorbid conditions discovered from the 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.

2021

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)

NLP in Healthcare



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** .

Pretrained NER Models

ner_ade_clinical	ner_posology_experimental
ner_posology_greedy	ner_drugs
ner_risk_factors	ner_deid_sd
jsl_ner_wip_clinical	ner_posology_large
ner_human_phenotype_gene_clinical	ner_deid_large
jsl_ner_wip_greedy_clinical	ner_posology
ner_cellular	ner_deidentify_dl
ner_cancer_genetics	ner_deid_enriched
jsl_ner_wip_modifier_clinical	ner_bacterial_species
ner_drugs_greedy	ner_drugs_large
ner_deid_sd_large	ner_clinical_large
ner_diseases	jsl_rd_ner_wip_greedy_clinical
nerdl_tumour_demo	ner_medmentions_coarse
ner_deid_subentity_augmented	ner_radiology_wip_clinical
ner_jsl_enriched	ner_clinical
ner_genetic_variants	ner_chemicals
ner_bionlp	ner_deid_synthetic
ner_measurements_clinical	ner_events_clinical
ner_diseases_large	ner_posology_small
ner_radiology	ner_anatomy_coarse
ner_deid_augmented	ner_human_phenotype_go_clinical
ner_anatomy	ner_jsl_slim
ner_chemprot_clinical	ner_jsl
	ner_jsl_greedy
	ner_events_admission_clinical

400+ entities from 100+ models

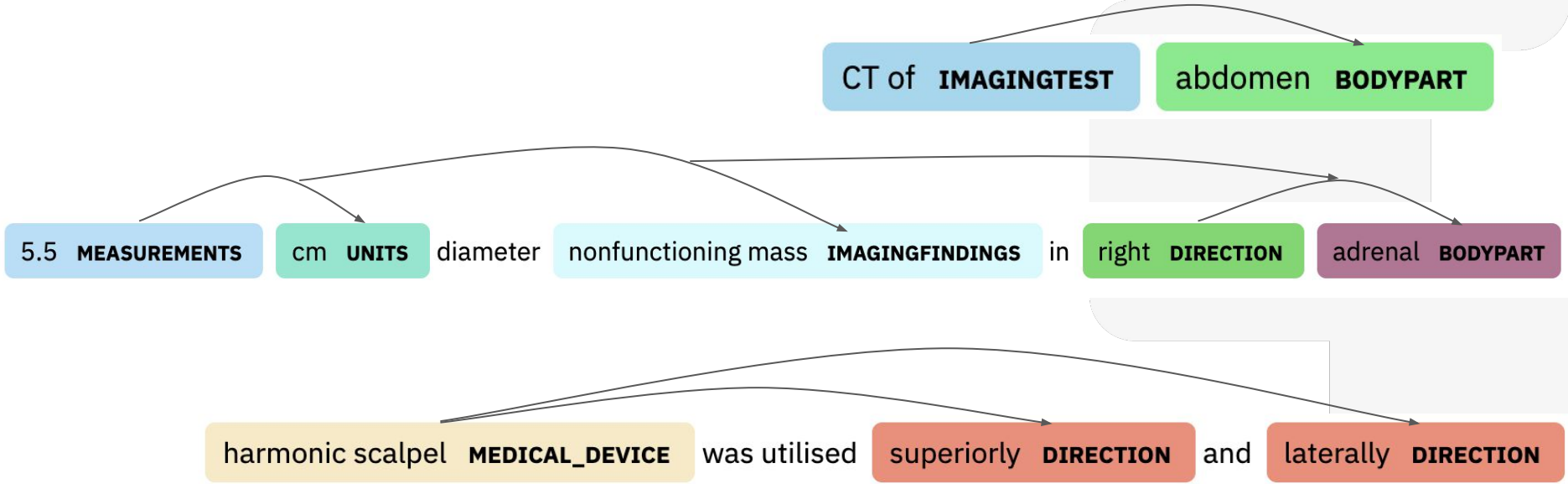
Assertion Status

"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

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."



Entity Resolution

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, UMLS, ATC, HPO,

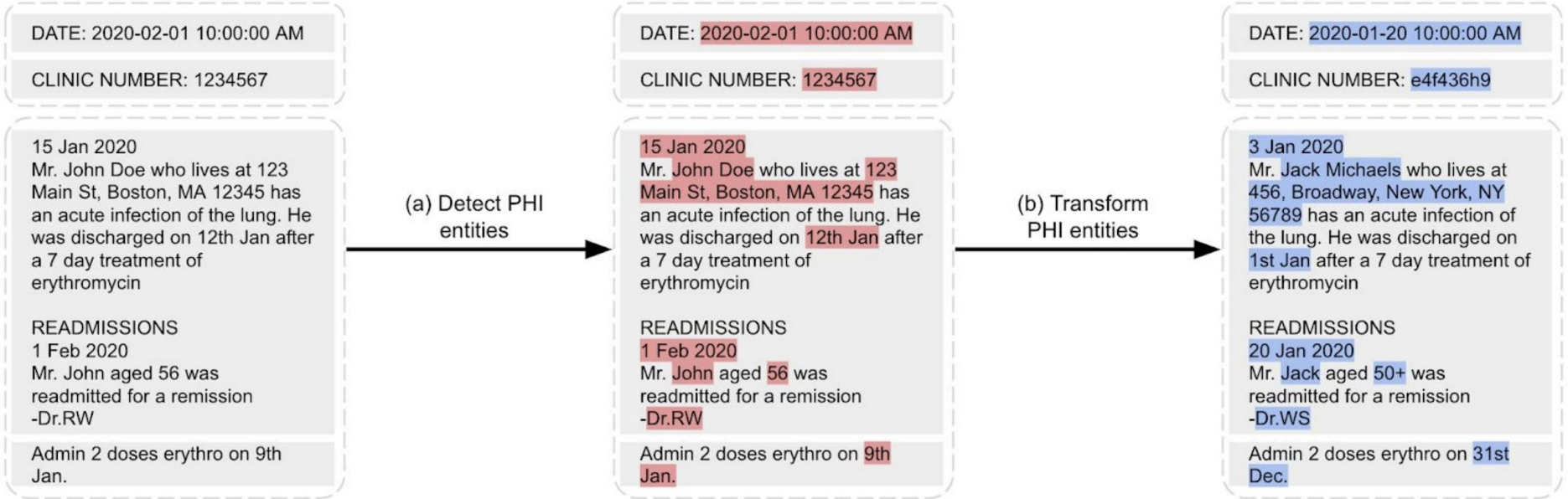
Term	Vocab	Code	Explanation (ground truth)
CT	CPT-4	76497	Unlisted computed tomography procedure
CT of abdomen	CPT-4	74150	Computed tomography, abdomen; without contrast material

weighted Sentence Chunk Embeddings (after 3.2.0)

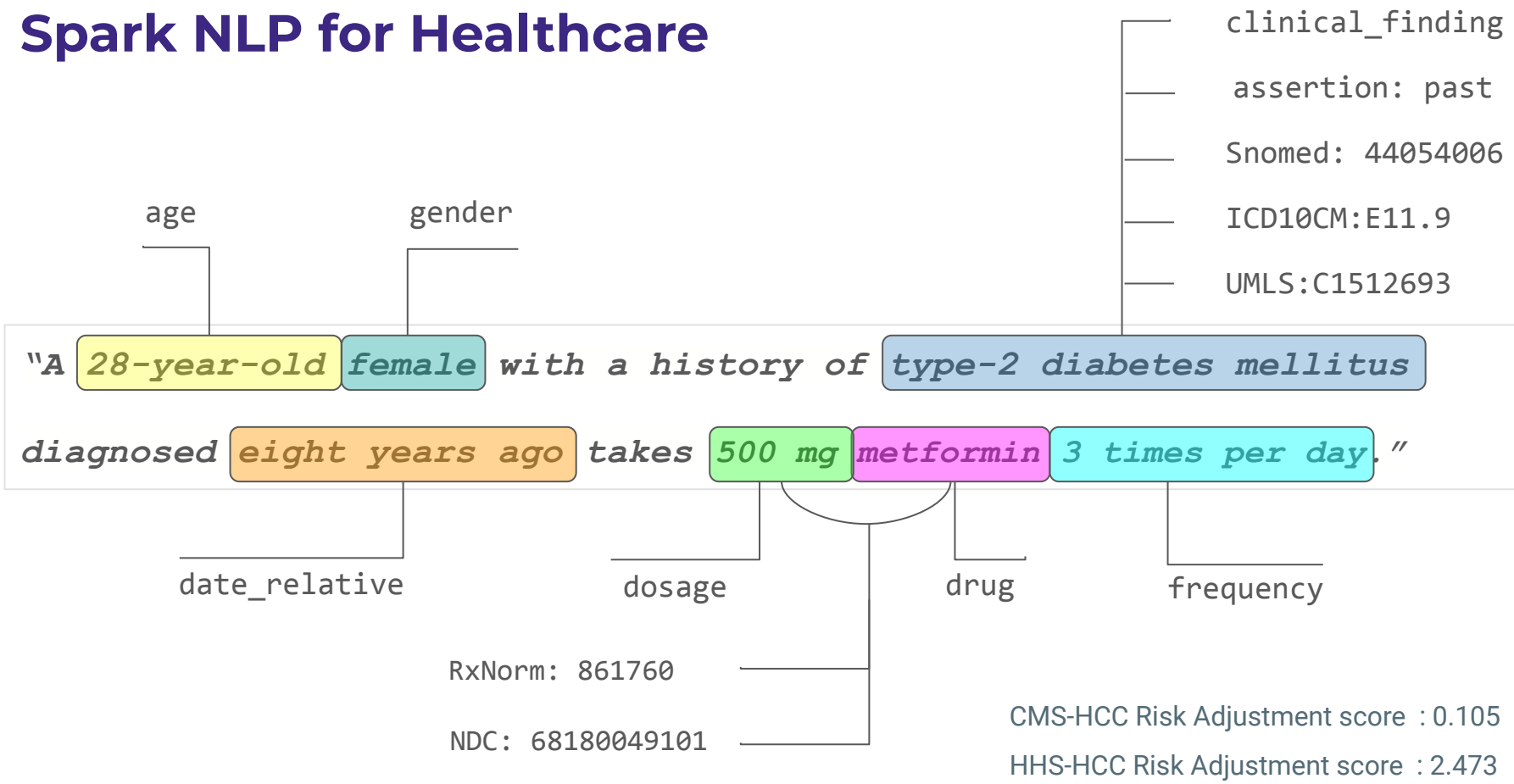
Term	Vocab	Code	Explanation (ground truth)
CT	CPT-4	74150	Computed tomography, abdomen; without contrast material

De-Identification

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

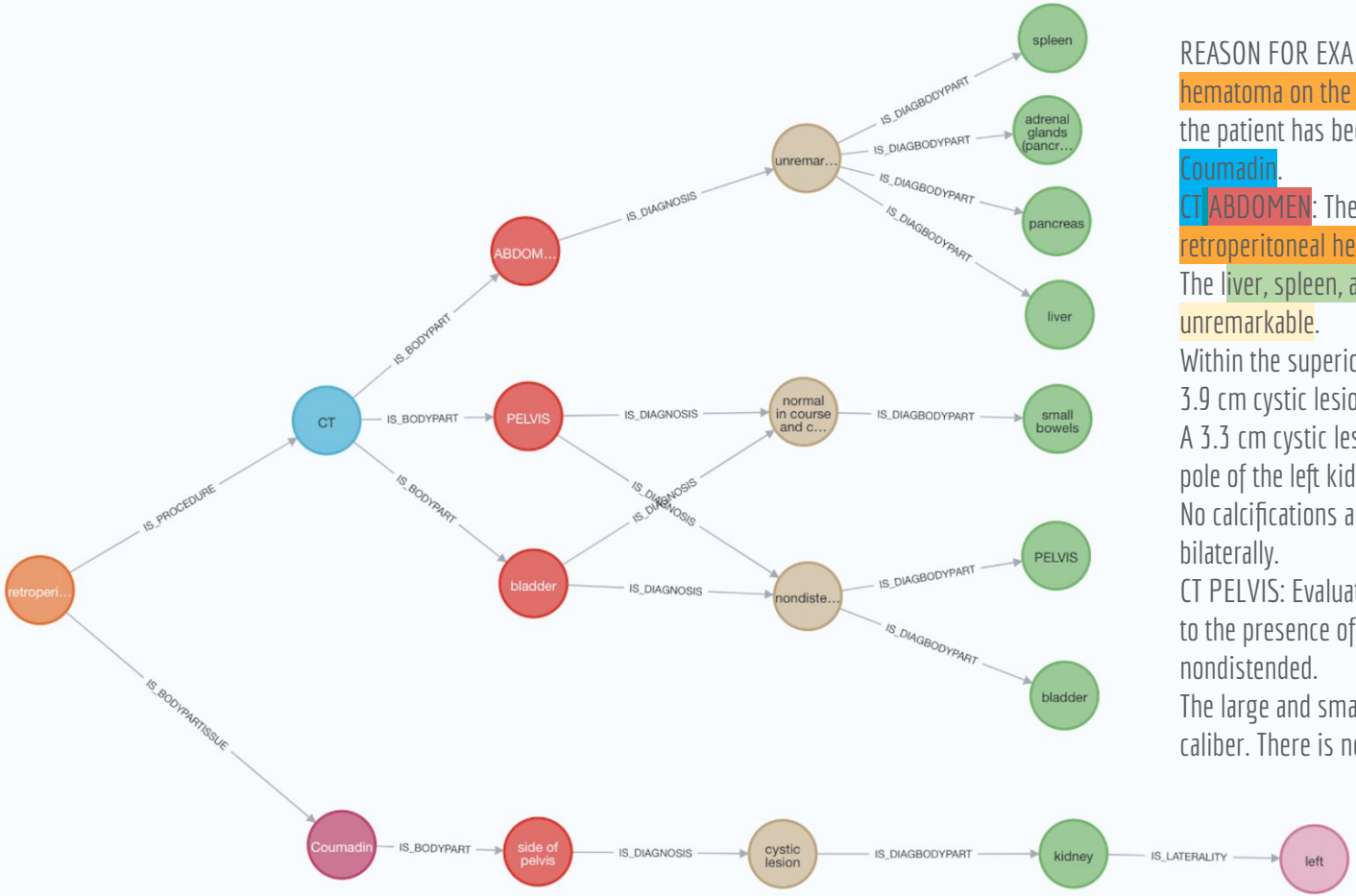


Spark NLP for Healthcare



*(20) Exam(1) Procedure(1) BodyPart(4) Diagnosis(4) DiagnosisBodyPart(8) BodyPartIssue(1) Laterality(1)

*(21) IS_BODYPARTISSUE(1) IS_PROCEDURE(1) IS_BODYPART(4) IS_DIAGNOSIS(6) IS_DIAGBODYPART(8) IS_LATERALITY(1)



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.

neo4j

Spark NLP for Healthcare

Named Entity Recognition

ICD10 Resolver

Snomed Resolver

UMLS Resolver

Assertion Status Detection

Risk Adj. Module

RxNorm Resolver

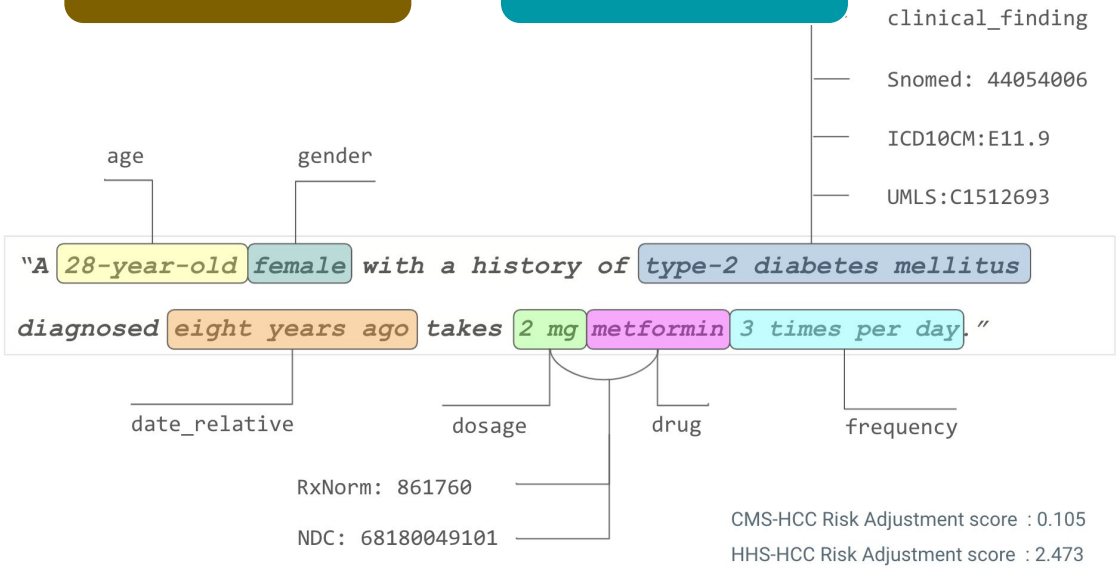
Relationship Extraction

Sentence Splitter

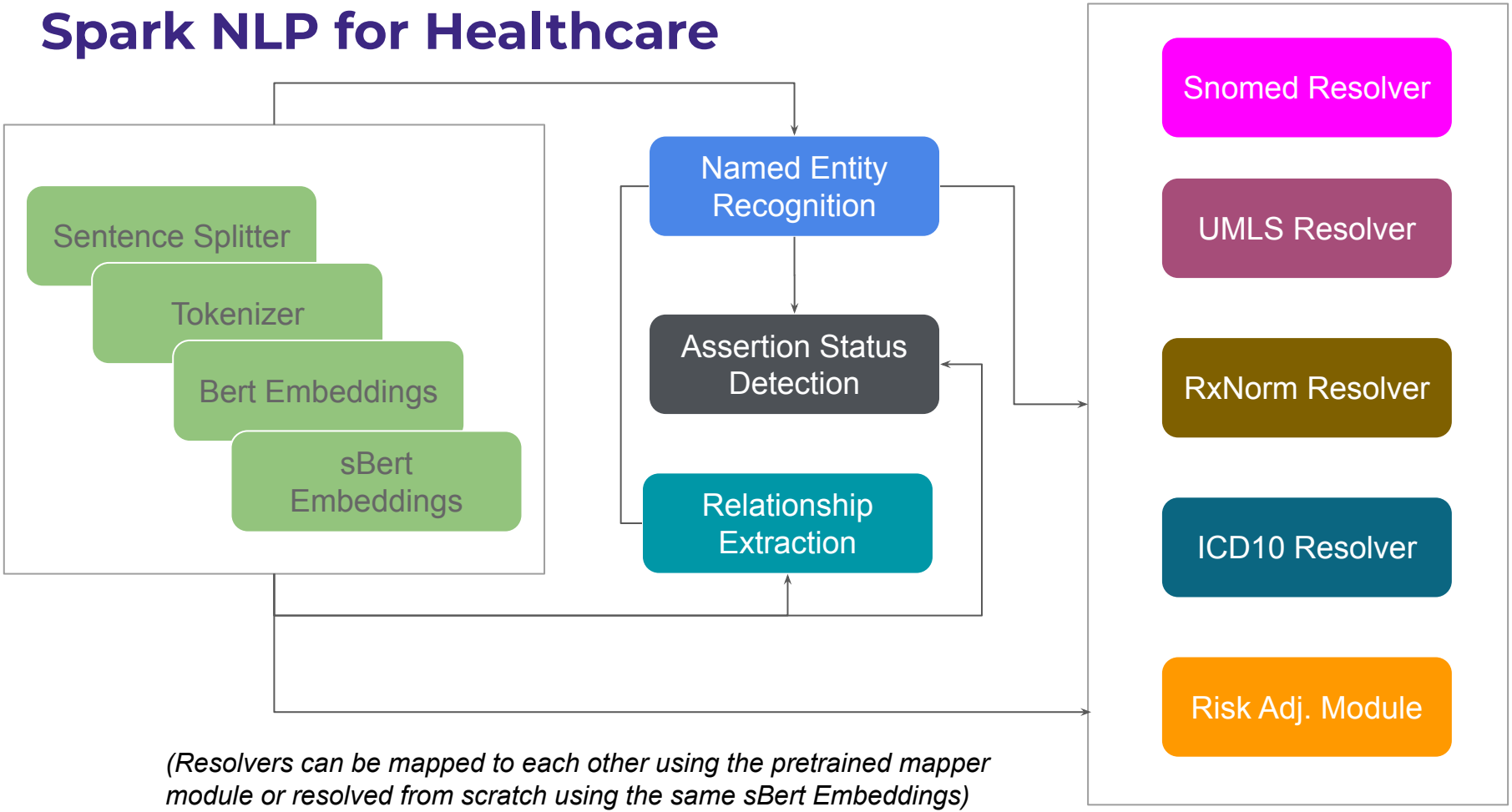
Tokenizer

Bert Embeddings

sBert Embeddings

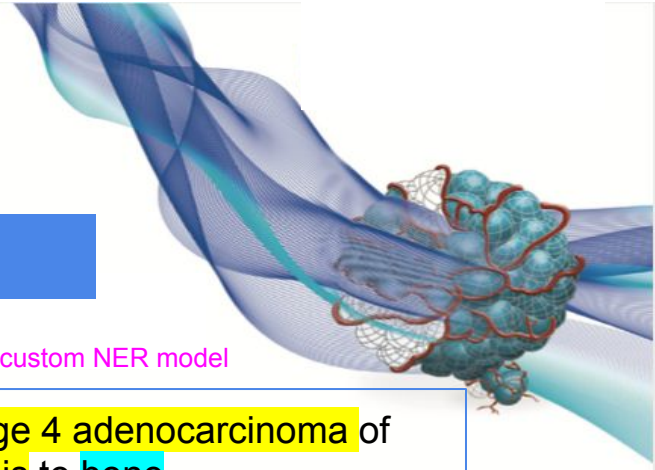


Spark NLP for Healthcare

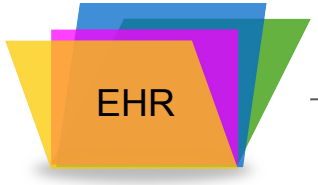


(Resolvers can be mapped to each other using the pretrained mapper module or resolved from scratch using the same sBert Embeddings)

NLP in Healthcare



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



Extract metastasis related sentences with mets-classifier

Annotate your own data and train a custom NER model

Recently diagnosed, stage 4 adenocarcinoma of both lungs with metastasis to bone.
CT scan shows no indication of mets on brain.

Extract named entities with Spark NLP NERDL and assign assertion statuses with AssertionDL model

Feature extraction & engineering

- *The frequency of clinical visits using document dates*
- *The number of positive site (organ) entities (hits by window)*
- *The number of Radiology/Oncology/Pathology reports in the last x days*
- *The number of tests applied in the last x days*
- *The number of diseases detected in the last x days*
- *Family history and social health determinants, etc.*
- *Date extraction and normalization*

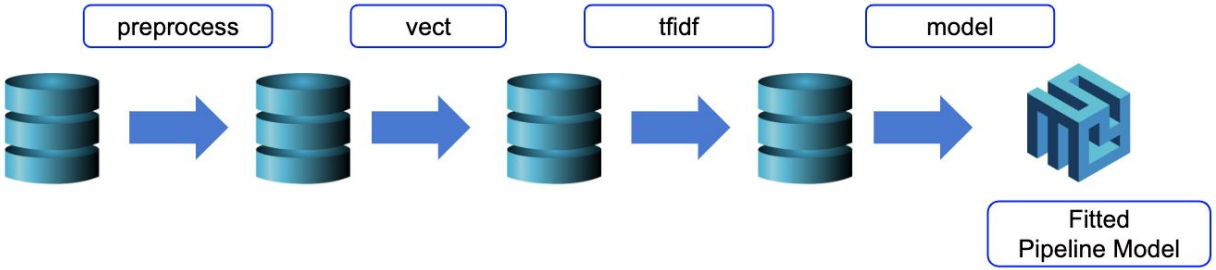
Text embeddings thru clinical word embeddings

Train a DL & ML Model

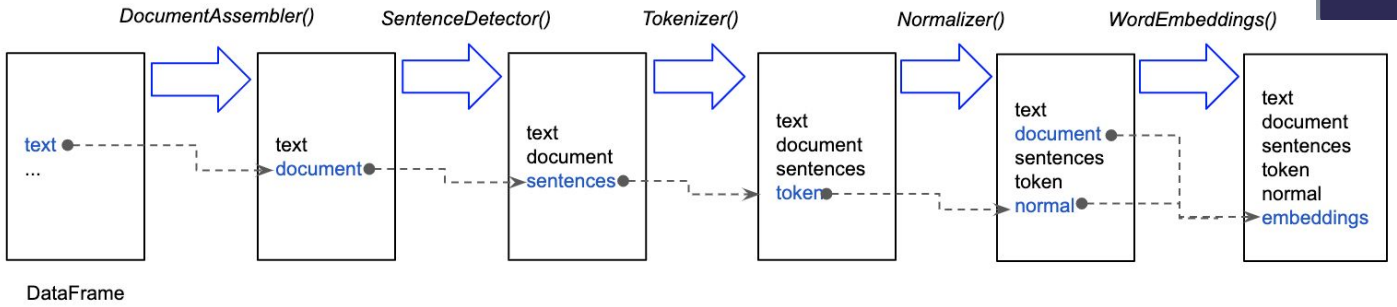
Prediction: Bone metastasis on June 2018

Spark NLP for Healthcare

Pipeline of annotators




```
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)
```

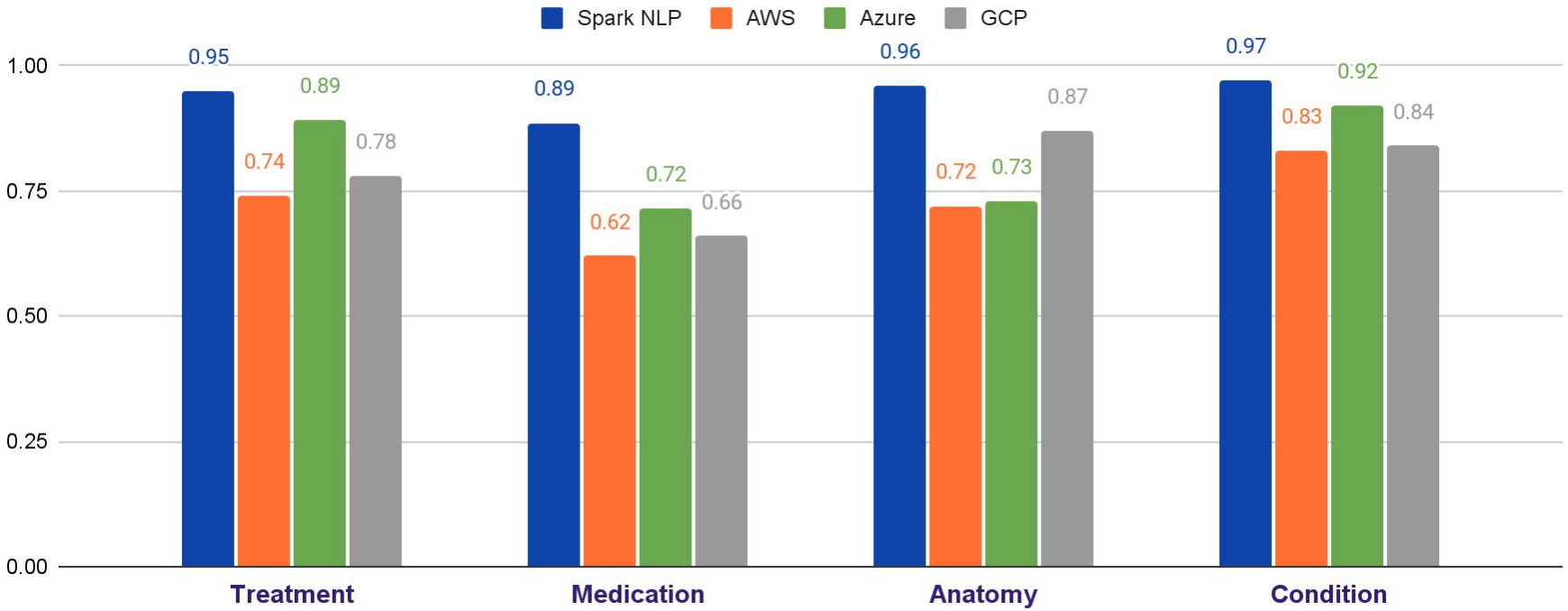


Attention (aka Bert) is all you need ?

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

* 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.

NER Benchmarks



~ 8K Sentences from MtSamples.com

Spark NLP for Healthcare - Deidentification

	English	German	French	Spanish	Italian
PATIENT	0.90	0.97	0.94	0.92	0.91
DOCTOR	0.94	0.98	0.99	0.92	0.92
HOSPITAL	0.91	1.00	0.94	0.86	0.90
DATE	0.98	1.00	0.98	0.99	0.98
AGE	0.94	0.99	0.86	0.98	0.98
PROFESSION	0.84	1.00	0.81	0.91	0.89
ORGANIZATION	0.77	0.94	0.77	0.83	0.74
STREET	0.9794	0.9802	0.8986	0.9448	0.9754
CITY	0.8331	0.9874	0.8643	0.8377	0.9678
COUNTRY	0.8083	0.9823	0.8983	0.8662	0.9262
PHONE	0.9412	0.882	0.9785	0.9027	0.9815
USERNAME	0.9215	1	0.9239	0.7407	0.9091
ZIP	0.9928	-	1	0.9895	0.9867

DATE: 2020-01-20 10:00:00 AM

CLINIC NUMBER: e4f436h9

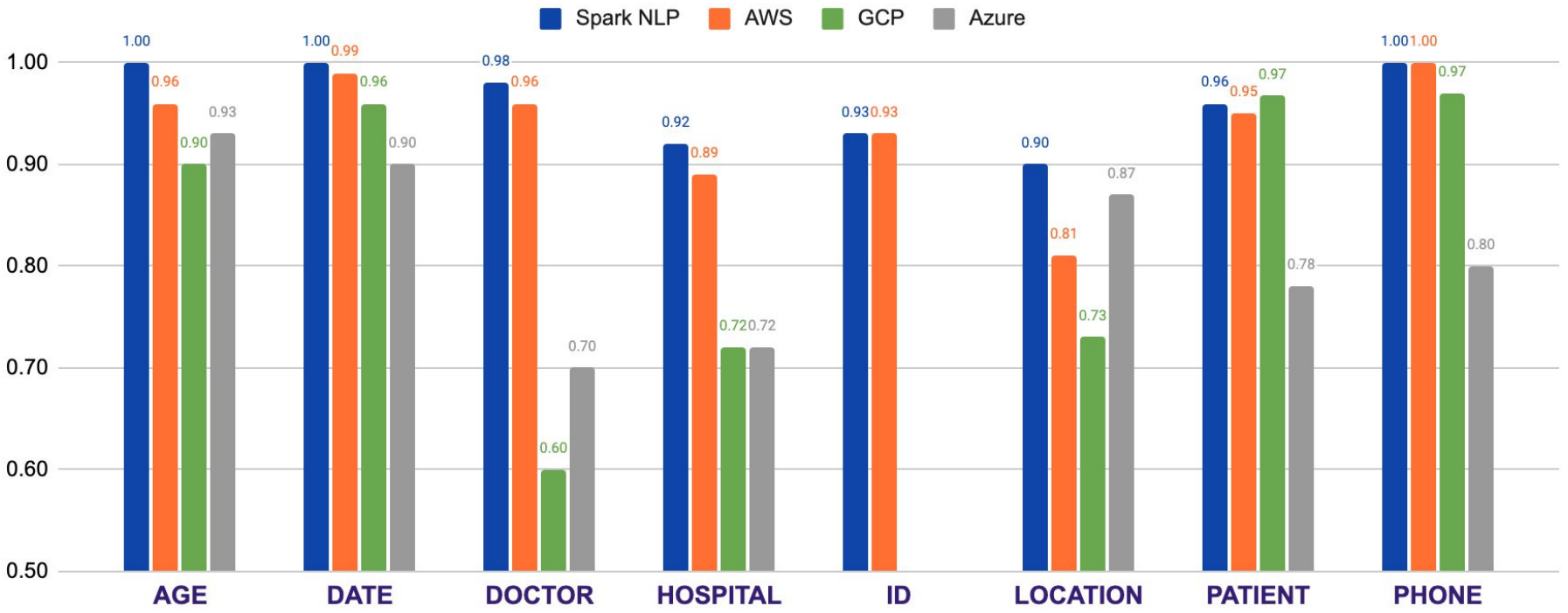
3 Jan 2020
Mr. Jack Michaels who lives at 456, Broadway, New York, NY 56789 has an acute infection of the lung. He was discharged on 1st Jan after a 7 day treatment of erythromycin

READMISSIONS
20 Jan 2020
Mr. Jack aged 50+ was readmitted for a remission
-Dr.WS

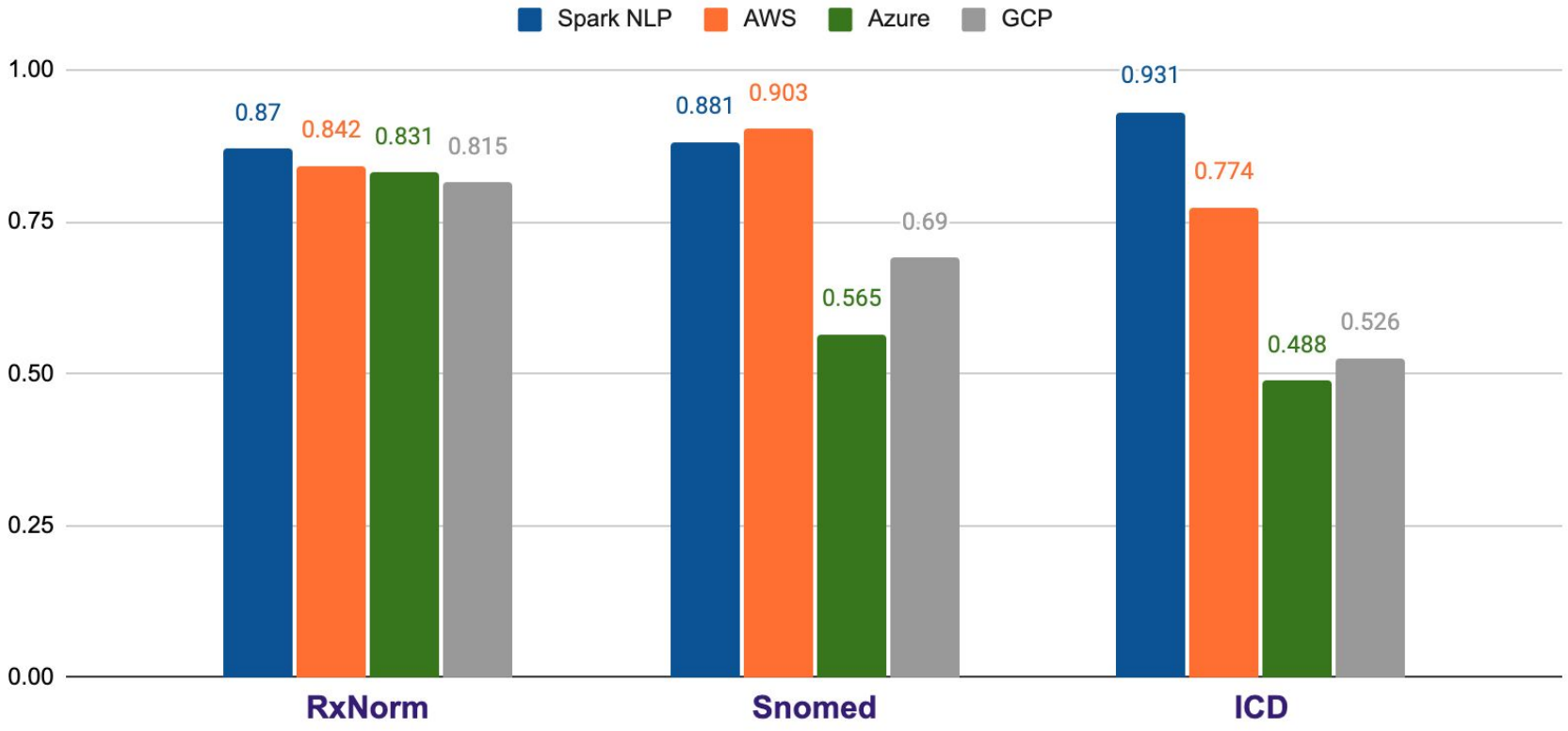
Admin 2 doses erythro on 31st Dec.

Sub-entity (13-entity)

Deidentification Benchmarks



Top - 5 Results



~ 8K Sentences from MtSamples.com

Thank you !

Veysel Kocaman
Principal Data Scientist
John Snow Labs