Automated Medical Data De-Identification and Obfuscation

NCI virtual workshop on Medical Image De-Identification (MIDI)
May 22, 2023

Jiri Dobes
Head of Solutions
John Snow Labs

John Snow LABS

John Snow Labs is the team behind Spark NLP

Most popular

NLP library in the enterprise

O'Reilly Media

59% share

of healthcare NLP teams use Spark NLP

Gradient Flow

#1 Accuracy

on 20 benchmarks in peer-reviewed papers

Papers with Code

HIPAA privacy rule: Need to de-identify data

No disclosure of PHI

- **§ 164.502 Uses and disclosures of protected health information:** General rules. (a) *Standard.* A covered entity or business associate **may not use** **or** **disclose** protected health information except as permitted…

Can disclose de-ID docs

- **§ 164.502(d)(2) Uses and disclosures of de-identified information.** Health information that meets the standard and implementation specifications for de-identification under § 164.514(a) and (b) is considered not to be individually identifiable health information, i.e., de-identified.
  - Cannot disclose how to re-identify or re-identify
  - No non-disclosure requirements

  The requirements of this subpart do not apply to information that has been de-identified in accordance with the applicable requirements of § 164.514, …

De-identification methods

- § 164.514(a) Information which does not identify and there is no reasonable basis to believe that can be used to identify individual is not IIHI
- § 164.514(b)
  - (1) **Apply generally accepted statistical methods** to determine… …**and** **documents** **the method and results**
  - (2) **Safe harbor** – 18 specified identifiers are removed.

John Snow Labs & De-identification capabilities

Healthcare NLP & Visual NLP

- Built for enterprise deployment
- Runs on-premise or cloud

Various formats and languages

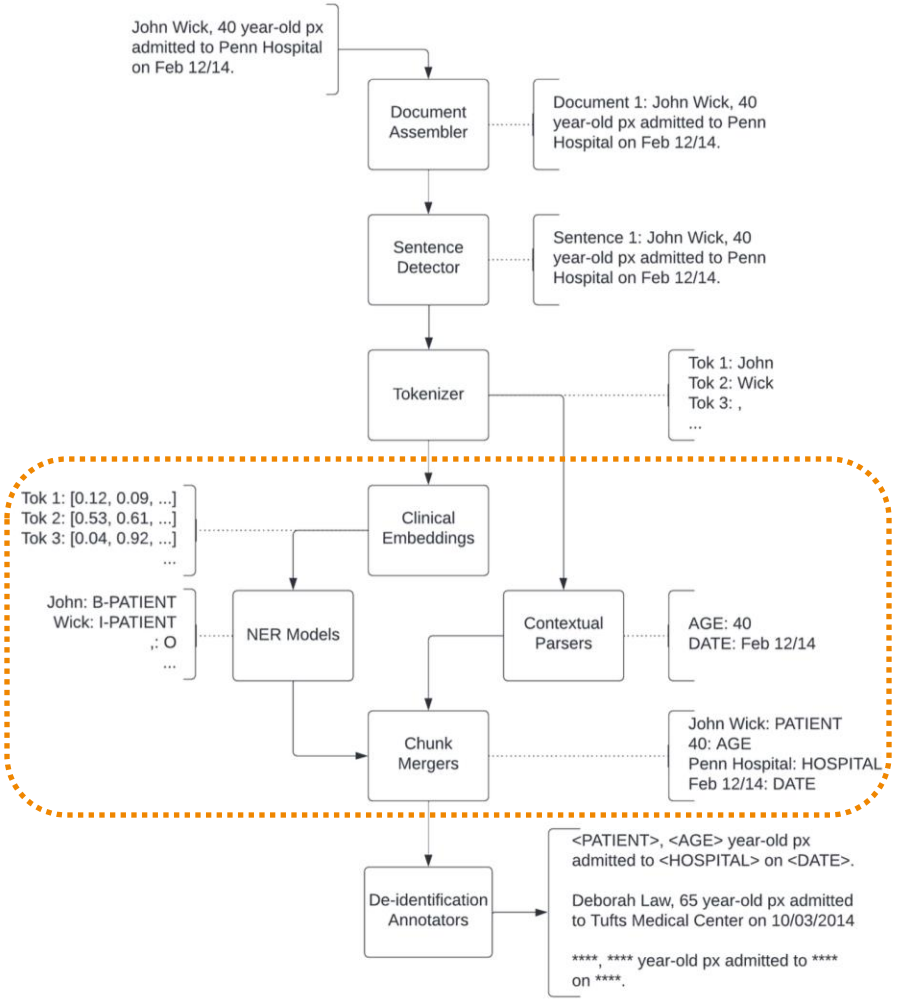
- Text formats: **unstructured txt**, xml, JSON
- Documents: **scan PDF**, text PDF, office format
- Images: **DICOM**, tiff, png, jpeg

- Languages (05/2023): EN, DE, FR, SP, IT, PT, RO.
  - Growing list based on customer demand.

Masking vs. obfuscation

- Masking:
  - John Smith → <NAME> or “*****”

- Obfuscation:
  - John Smith → Rob Davis
  - 5/22/2023 → 7/12/2022



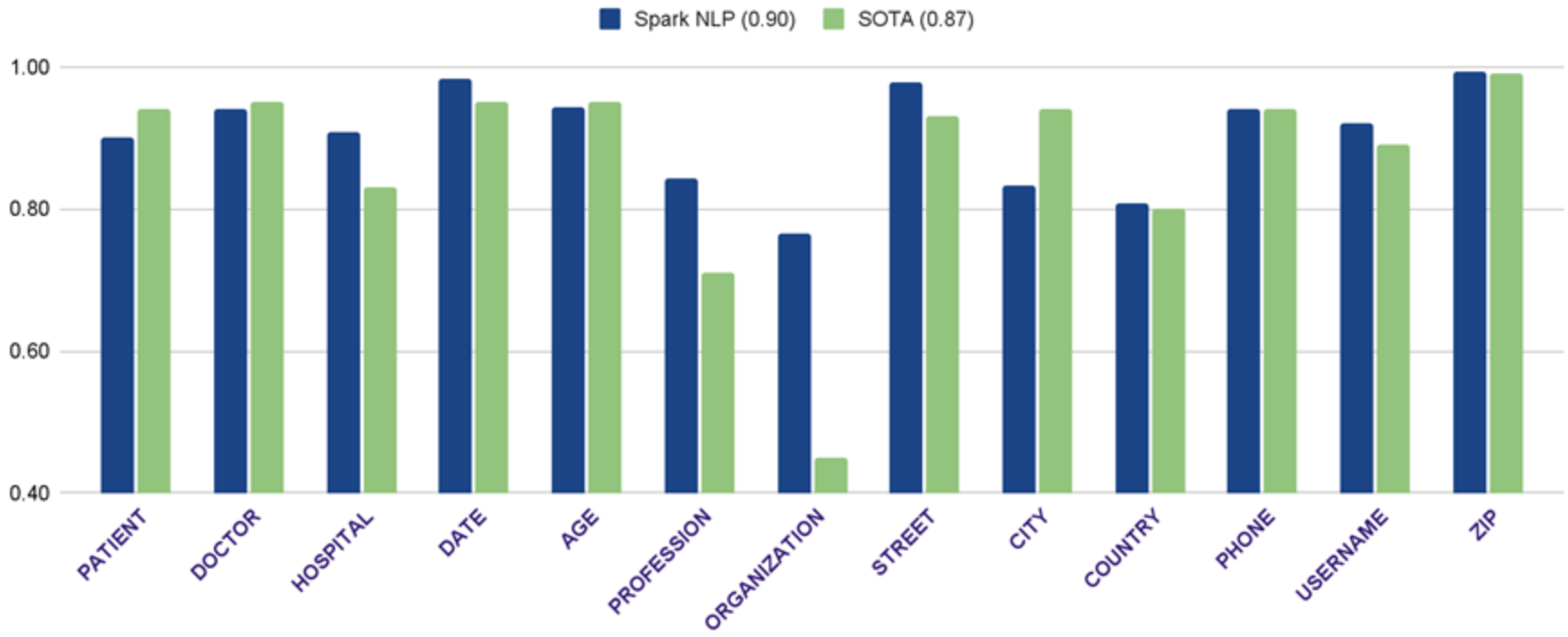

```
documentAssembler = nlp.DocumentAssembler()\
    .setInputCol("text")\
    .setOutputCol("document")

# Sentence Detector annotator, processes various sentences per line
sentenceDetector = nlp.SentenceDetector()\
    .setInputCols(["document"])\
    .setOutputCol("sentence")

# Tokenizer splits words in a relevant format for NLP
tokenizer = nlp.Tokenizer()\
    .setInputCols(["sentence"])\
    .setOutputCol("token")

# Clinical word embeddings trained on PubMED dataset
word_embeddings = nlp.WordEmbeddingsModel.pretrained("embeddings_clinical", "en", "clinical/models")\
    .setInputCols(["sentence", "token"])\
    .setOutputCol("embeddings")

# NER model trained on n2c2 (de-identification and Heart Disease Risk Factors Challenge) datasets)
clinical_ner = medical.NerModel.pretrained("ner_deid_generic_augmented", "en", "clinical/models") \
    .setInputCols(["sentence", "token", "embeddings"]) \
    .setOutputCol("ner")

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

deidentification = medical.DeIdentification() \
      .setInputCols(["sentence", "token", "ner_chunk"]) \
      .setOutputCol("deidentified") \
      .setMode("mask")\
      .setReturnEntityMappings(True)
      #.setMappingsColumn("MappingCol")

deidPipeline = nlp.Pipeline(stages=[
      documentAssembler,
      sentenceDetector,
      tokenizer,
      word_embeddings,
      clinical_ner,
      ner_converter,
      deidentification])
```

Enriching de-identification pipeline with regex and contextual parser

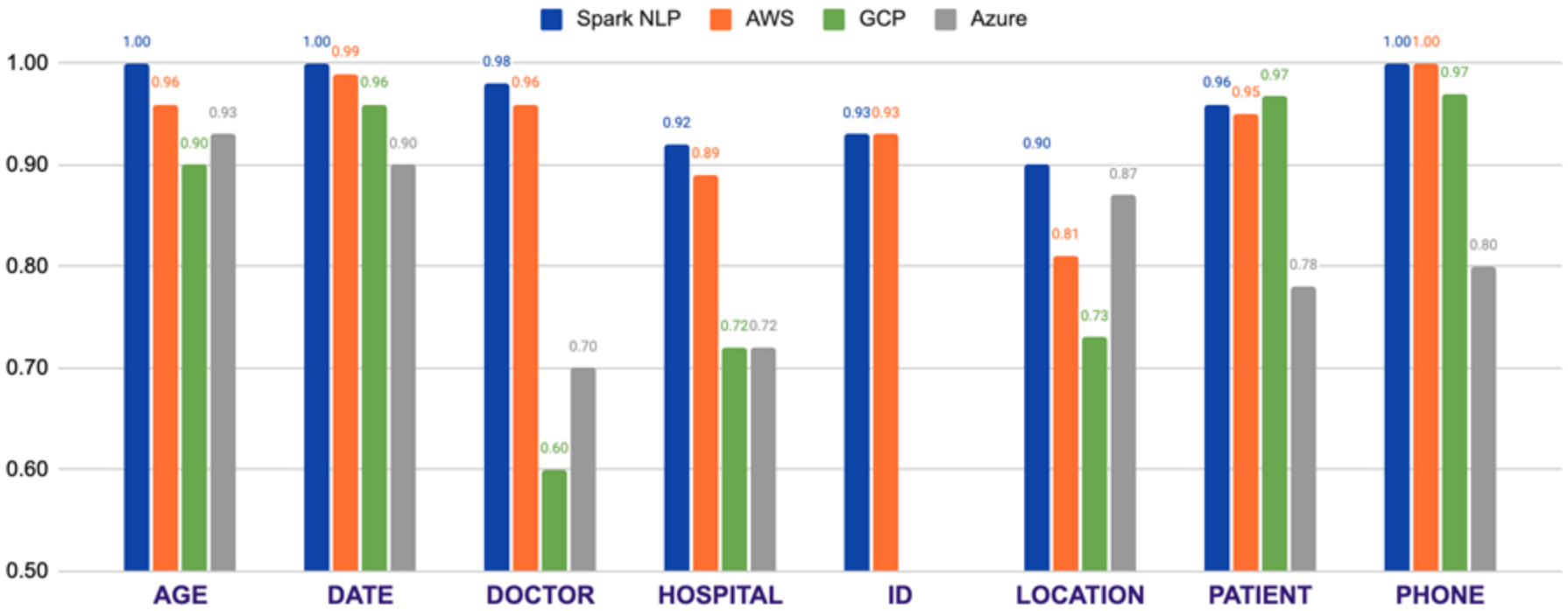
	English		German		Spanish		Portuguese		Italian		French	
Entity	NER	Pipeline	NER	Pipeline	NER	Pipeline	NER	Pipeline	NER	Pipeline	NER	Pipeline
Age	0.910	0.967	0.944	0.965	0.971	0.987	0.963	0.984	0.969	0.984	0.933	0.978
Date	0.973	0.988	0.999	0.999	0.965	0.978	0.989	0.995	0.985	0.986	0.991	0.997
ID	0.930	0.974	0.974	0.984	0.978	0.994	0.978	0.996	0.980	0.988	0.966	0.983
Location	0.803	0.927	0.797	0.855	0.870	0.903	0.958	0.968	0.971	0.985	0.868	0.956
Avg.	**0.904**	**0.964**	**0.929**	**0.951**	**0.946**	**0.965**	**0.972**	**0.986**	**0.976**	**0.986**	**0.939**	**0.979**
PHI	0.948	0.982	0.958	0.966	0.974	0.983	0.992	0.994	0.984	0.992	0.986	0.996

- When de-identification pipeline is enriched with regex and contextual parser (not just NERs), an average improvement is around 10% across all entities.
- The most drastic improvements occurred in the Location and Age entities, with improvements of 12% and 5% respectively.
- When it comes to binary PHI recognition performance, the gain was between 1 and 4%, exceeding 95% accuracy in all the languages supported, even exceeding 99% in some of the languages

Deidentification Benchmarks



De-Identification Benchmarks (en)



Source: https://github.com/JohnSnowLabs/spark-nlp-workshop/blob/master/tutorials/Certification_Trainings/Healthcare/4.3.Clinical_Deidentification_SparkNLP_vs_Cloud_Providers_Comparison.ipynb

Obfuscation Consistency

Name consistency: Mapping *Jane Doe* and *Jane* to the same first name. If we map *Jane Doe* to a fake name (e.g. *Nancy Smith*), the next name entity (*Jane*) that corresponds to the same patient should also be replaced by *Nancy*. In addition, when there are multiple clinical notes for the same patient, the same mapping should be made across different documents to have a consistent obfuscation for several concerns such as traceability or regulations. Hence, this mapping can be aligned to patient IDs so that every patient will get different mapping even for the same names (e.g. "Jane" will be mapped to "Mary" for patient-1 whereas the same name is mapped to "Jen" for patient-2).

Gender consistency: Mapping *Jane* to a feminine American name (or a feminine British name if needed).

Age consistency: Specifying a proper age range (i.e. age groups such as 5-12 years for children, 20-39 years for adults etc.) to make the obfuscation within that age group. The age obfuscation should be consistent here due to some phrases (e.g. *lady*, *lovely*) that hint to an adult lady. Hence we should replace *78* with a reasonable age (e.g. *40* but not *5* or *12*).

Clinical consistency: Note that *Jane* needs to "remain female" also because she has a history of breast cancer.

Day shift consistency: Shifting the days based on a predefined list of shift values per patient ID (e.g. plus 2 days for patient-1, minus 5 days for patient-2 etc.) as well as allowing a completely random shift given a range.

Date format consistency: If *April 2020* needs to be shifted by a random number of days, then the result should be in the same format (i.e. *March 2020* and not *3/3/2020*). Moreover, since there is no day information in the original date entity and it is not in a proper date format, this date should be normalized to a proper date format (e.g. *04/1/2020*) at first in order to apply a day shift.

Length consistency: In order to keep the length of the original text intact, it's often required to replace the selected entities with the same length of fake entities. If same length is not possible, adding or deleting characters can force it into the same length.

Presented by John Snow LABS

Thank you.

Spark NLP for Healthcare - Deidentification

	English	German	French	Spanish	Italian	Portuguese	Romanian
PATIENT	0.9	0.97	0.94	0.92	0.91	0.95	0.87
DOCTOR	0.94	0.98	0.99	0.92	0.92	0.93	0.96
HOSPITAL	0.91	1.00	0.94	0.86	0.90	0.90	0.8
DATE	0.98	1.00	0.98	0.99	0.98	0.98	0.91
AGE	0.94	0.99	0.86	0.98	0.98	0.98	0.97
PROFESSION	0.84	1.00	0.81	0.91	0.89	0.90	0.83
ORGANIZATION	0.77	0.94	0.77	0.83	0.74	0.97	0.37
STREET	0.98	0.98	0.90	0.94	0.98	1.00	0.99
CITY	0.83	0.99	0.86	0.84	0.97	0.98	0.96
COUNTRY	0.81	0.98	0.90	0.87	0.93	0.91	0.82
PHONE	0.94	0.88	0.98	0.90	0.98	0.99	0.98
USERNAME	0.92	1.00	0.92	0.74	0.91	0.88	-
ZIP	0.99	-	1.00	0.99	0.99	0.99	0.98

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.

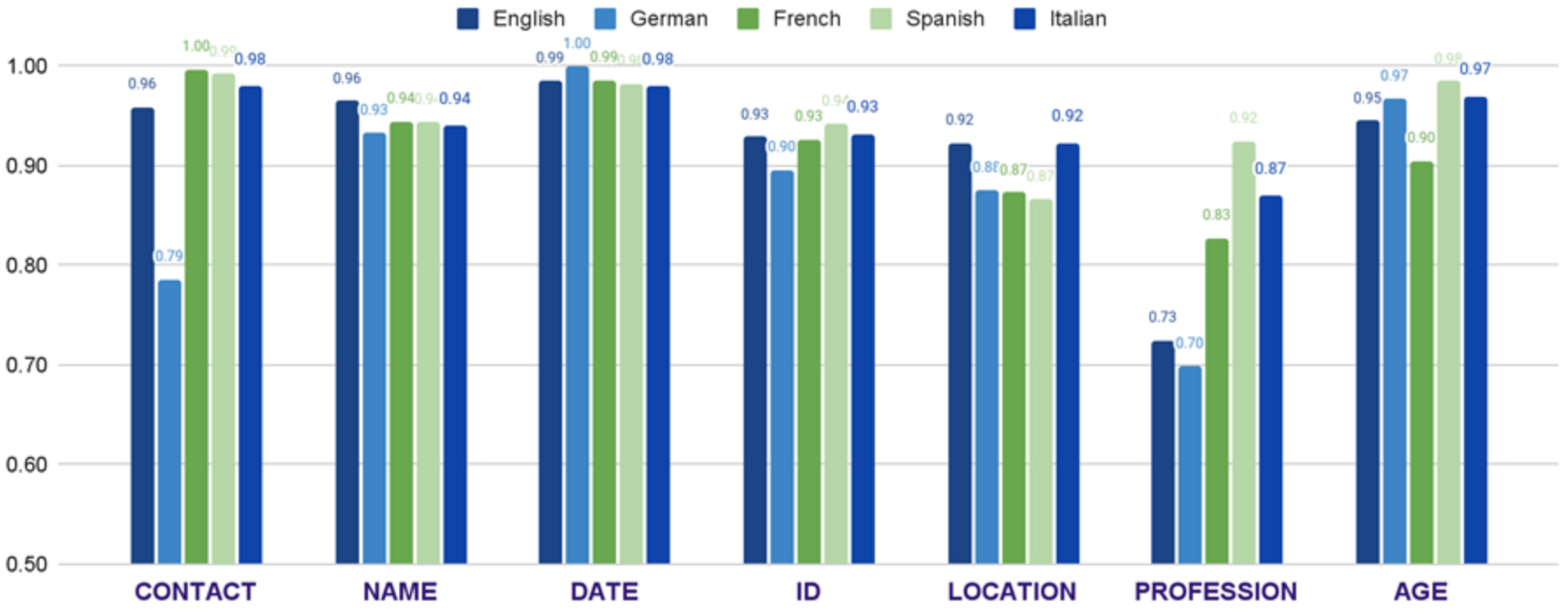
Sub-entities (13-entity)

`MEDICALRECORD` , `ORGANIZATION` , `DOCTOR` , `USERNAME` , `PROFESSION` , `HEALTHPLAN` , `URL` , `CITY` , `DATE` , `LOCATION-OTHER` , `STATE` , `PATIENT` , `DEVICE` , `COUNTRY` , `ZIP` , `PHONE` , `HOSPITAL` , `EMAIL` , `IDNUM` , `SREET` , `BIOID` , `FAX` , `AGE`

Generic entities (7-entity)

`DATE` , `NAME` , `LOCATION` , `PROFESSION` , `CONTACT` , `AGE` , `ID`

Deidentification Benchmarks



Generic (7-entity)

Out of the box de-ID entities

- ACCOUNT, AGE, BIOID, CITY, CONTACT, COUNTRY, DATE, DEVICE, DLN, DOCTOR, EMAIL, FAX, HEALTHPLAN, HOSPITAL, ID, IDNUM, IPADDR, LICENSE, LOCATION, LOCATION-OTHER, MEDICALRECORD, NAME, ORGANIZATION, PATIENT, PHONE, PLATE, PROFESSION, SSN, STREET, STATE, URL, USERNAME, VIN, ZIP