Automated Medical Data De-Identification and Obfuscation

Veysel Kocaman
Head of Data Science
John Snow Labs

Jiri Dobes
Head of Solutions
John Snow Labs

Agenda

- Motivation & background
- John Snow Labs' de-identification capabilities
- Technical solution & benchmarks

Agenda

- **Motivation & background**
- John Snow Labs’ de-identification capabilities
- Technical solution & benchmarks

Increasing concerns but also improvements in NLP

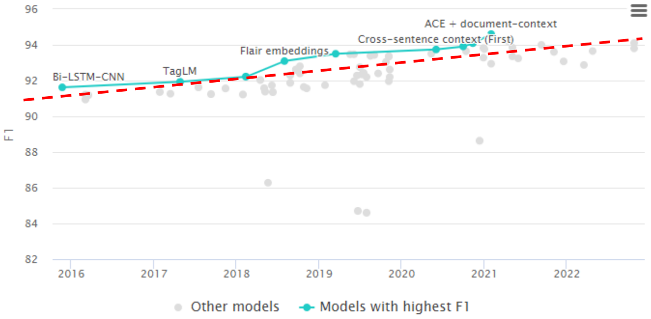
Increasing privacy concerns

- HIPAA privacy rule (2003)
- GDPR (2016)
- Other regions – Canada, Australia…

Increasing volume of data[1])

- 30% of the world's data volume is being generated by the healthcare industry
- 36% CAGR data grow in Healthcare
  - 6% higher growth than manufacturing and 10% higher than financial services

Improving capabilities of NLP[2])



- NER – 0.5% error decrease per year
- Already surpassed human performance on de-identification task

Sources: 1. D. Reinsel, J. Gantz, J. Rydning: The Digitization of the World. From Edge to Core. An IDC White Paper. November 2018.
2. Paper w ith code – benchmark Named Entity Recognition on CoNLL 2003 (English).

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 **<u>may not use or disclose</u>** 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.

…similar situation in Europe

GDPR: No disclosure of PHI

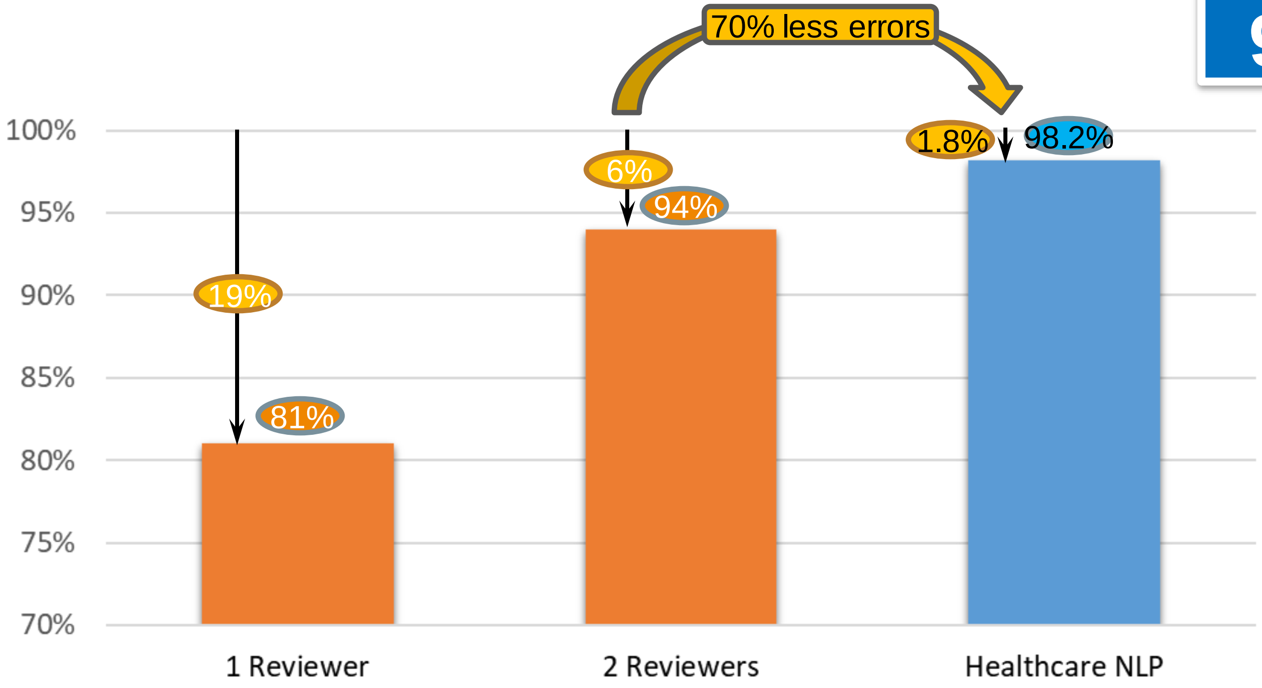
No protection for anonymous information

(26) The principles of data protection should apply to any information concerning an identified or identifiable natural person. Personal data which have undergone pseudonymisation, which could be attributed to a natural person by the use of additional information should be considered to be information on an identifiable natural person. **To determine whether a natural person is identifiable, account should be taken of all the means reasonably likely to be used**, such as singling out, either by the controller or by another person to identify the natural person **directly or indirectly**. To ascertain whether means are reasonably likely to be used to identify the natural person, account should be taken of all objective factors, such as the **costs of and the amount of time required for identification, taking into consideration the available technology** at the time of the processing and technological developments. The principles of data protection should therefore not apply to anonymous information, namely information which does not relate to an identified or identifiable natural person or to **personal data rendered anonymous in such a manner that the data subject is not or no longer identifiable**. This Regulation does not therefore concern the processing of such anonymous information, including for statistical or research purposes.

NLP: ~70% less error rate than manual effort

…and manual de-identification may not be as accurate as you believe.

Why not 99.xx%?



M. Douglass, G. D. Clifford, A. Reisner, G. B. Moody and Mark RG, "Computer-assisted de-identification of free text in the MIMIC II database," Computers in Cardiology, 2004, 2004, pp. 341-344, doi: 10.1109/CIC.2004.1442942

Agenda

- Motivation & background
- **John Snow Labs’ de-identification capabilities**
- Technical solution & benchmarks

John Snow Labs de-identification capabilities

Input format

PDF Searchable & scanned

TXT

DICOM

Digital Imaging and Communications in Medicine

Pixel level & embedded data

Office

W X O P

CSV

{JSON}

TIFF

PNG

JPEG

Languages

🇺🇸🇬🇧 🇫🇷

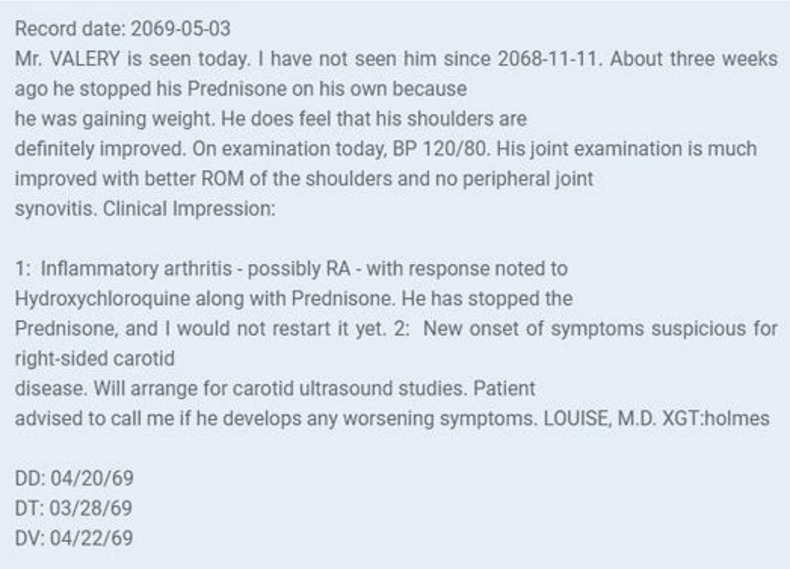
🇪🇸 🇮🇹

🇩🇪 🇷🇴

🇵🇹

- Growing number of languages based on customer's demand

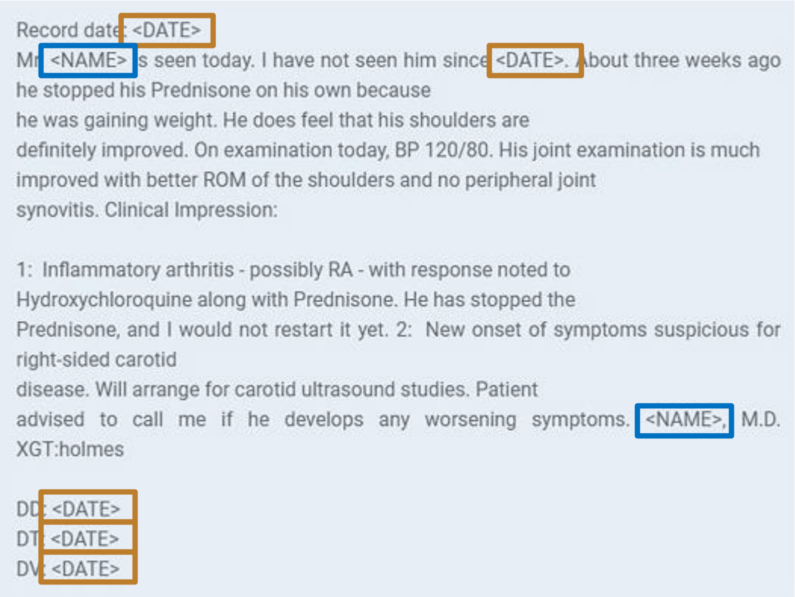
De-identification: **Masking**



Record date: 2069-05-03
Mr. VALERY is seen today. I have not seen him since 2068-11-11. About three weeks ago he stopped his Prednisone on his own because
he was gaining weight. He does feel that his shoulders are
definitely improved. On examination today, BP 120/80. His joint examination is much improved with better ROM of the shoulders and no peripheral joint
synovitis. Clinical Impression:

1: Inflammatory arthritis - possibly RA - with response noted to
Hydroxychloroquine along with Prednisone. He has stopped the
Prednisone, and I would not restart it yet. 2: New onset of symptoms suspicious for right-sided carotid
disease. Will arrange for carotid ultrasound studies. Patient
advised to call me if he develops any worsening symptoms. LOUISE, M.D. XGT:holmes

DD: 04/20/69
DT: 03/28/69
DV: 04/22/69



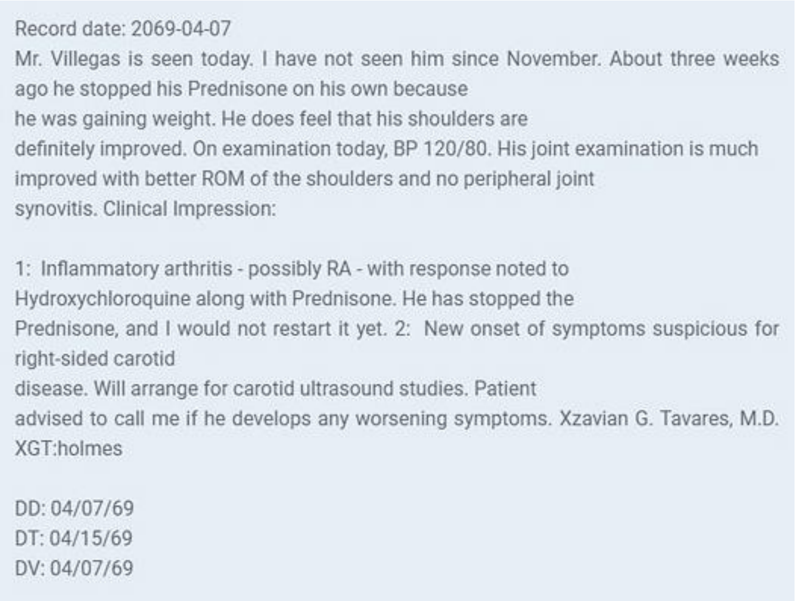
Record date: <DATE>
Mr. <NAME> is seen today. I have not seen him since <DATE>. About three weeks ago he stopped his Prednisone on his own because
he was gaining weight. He does feel that his shoulders are
definitely improved. On examination today, BP 120/80. His joint examination is much improved with better ROM of the shoulders and no peripheral joint
synovitis. Clinical Impression:

1: Inflammatory arthritis - possibly RA - with response noted to
Hydroxychloroquine along with Prednisone. He has stopped the
Prednisone, and I would not restart it yet. 2: New onset of symptoms suspicious for right-sided carotid
disease. Will arrange for carotid ultrasound studies. Patient
advised to call me if he develops any worsening symptoms. <NAME>, M.D. XGT:holmes

DD: <DATE>
DT: <DATE>
DV: <DATE>

- <NAME>, <DATE>, …
- ****

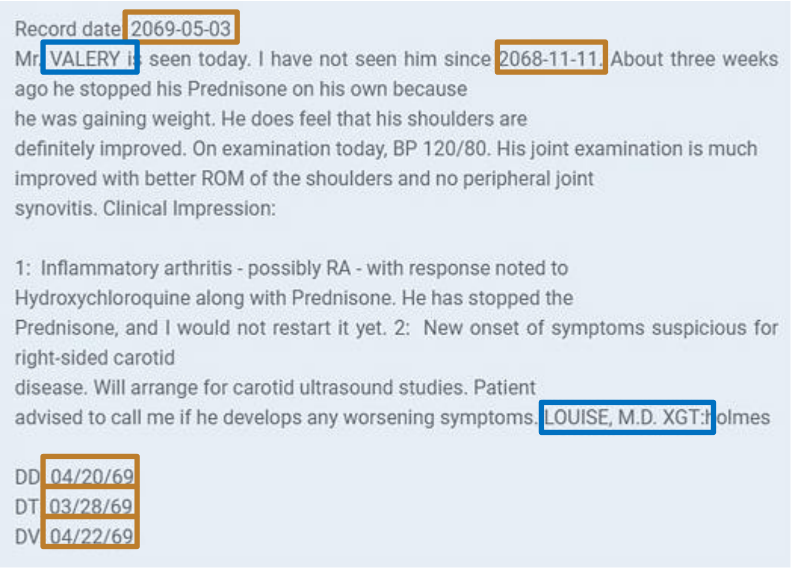
De-identification: **Obfuscation & faking**



Record date: 2069-04-07
Mr. Villegas is seen today. I have not seen him since November. About three weeks ago he stopped his Prednisone on his own because he was gaining weight. He does feel that his shoulders are definitely improved. On examination today, BP 120/80. His joint examination is much improved with better ROM of the shoulders and no peripheral joint synovitis. Clinical Impression:

1: Inflammatory arthritis - possibly RA - with response noted to Hydroxychloroquine along with Prednisone. He has stopped the Prednisone, and I would not restart it yet. 2: New onset of symptoms suspicious for right-sided carotid disease. Will arrange for carotid ultrasound studies. Patient advised to call me if he develops any worsening symptoms. Xzavian G. Tavares, M.D. XGT:holmes

DD: 04/07/69
DT: 04/15/69
DV: 04/07/69



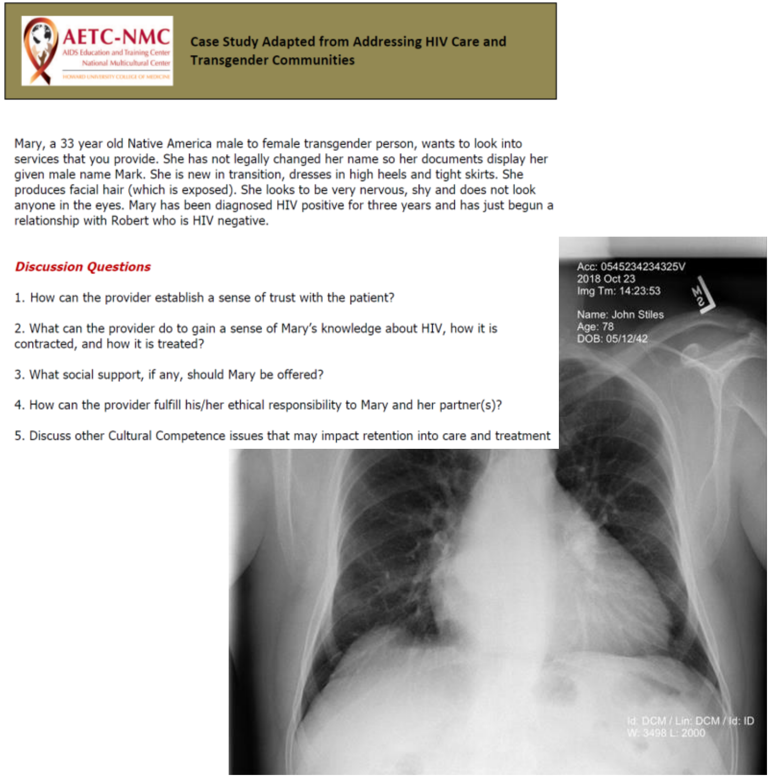
Record date 2069-05-03
Mr. VALERY is seen today. I have not seen him since 2068-11-11. About three weeks ago he stopped his Prednisone on his own because he was gaining weight. He does feel that his shoulders are definitely improved. On examination today, BP 120/80. His joint examination is much improved with better ROM of the shoulders and no peripheral joint synovitis. Clinical Impression:

1: Inflammatory arthritis - possibly RA - with response noted to Hydroxychloroquine along with Prednisone. He has stopped the Prednisone, and I would not restart it yet. 2: New onset of symptoms suspicious for right-sided carotid disease. Will arrange for carotid ultrasound studies. Patient advised to call me if he develops any worsening symptoms. LOUISE, M.D. XGT:holmes

DD 04/20/69
DT 03/28/69
DV 04/22/69

- Random name from dictionary
- Random shift of date

De-identification: Document masking



Case Study Adapted from Addressing HIV Care and Transgender Communities

Mary, a 33 year old Native America male to female transgender person, wants to look into services that you provide. She has not legally changed her name so her documents display her given male name Mark. She is new in transition, dresses in high heels and tight skirts. She produces facial hair (which is exposed). She looks to be very nervous, shy and does not look anyone in the eyes. Mary has been diagnosed HIV positive for three years and has just begun a relationship with Robert who is HIV negative.

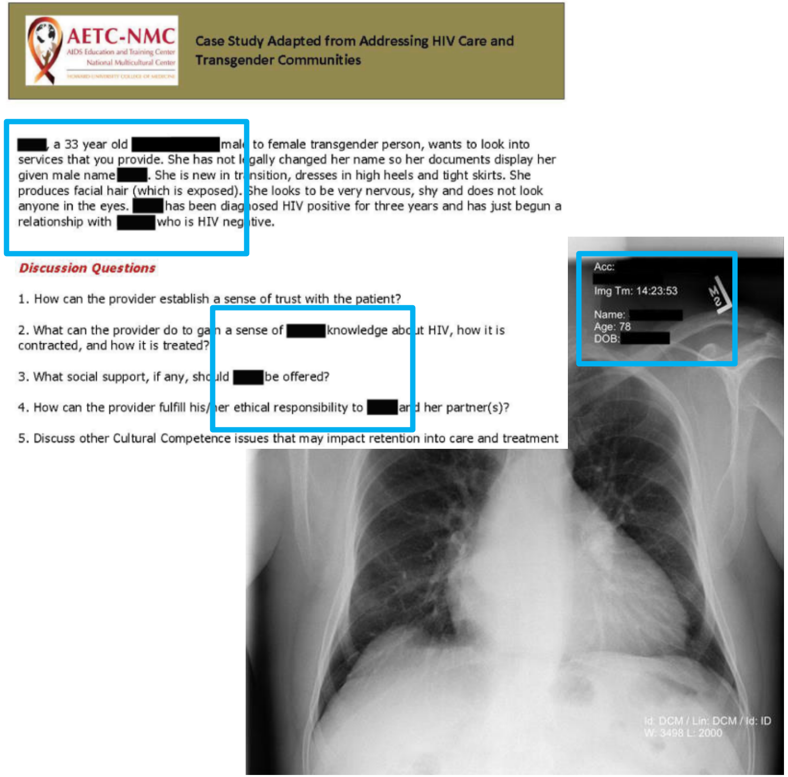
Discussion Questions

1. How can the provider establish a sense of trust with the patient?
2. What can the provider do to gain a sense of Mary's knowledge about HIV, how it is contracted, and how it is treated?
3. What social support, if any, should Mary be offered?
4. How can the provider fulfill his/her ethical responsibility to Mary and her partner(s)?
5. Discuss other Cultural Competence issues that may impact retention into care and treatment

Acc: 0545234234325V
2018 Oct 23
Img Tm: 14:23:53

Name: John Stiles
Age: 78
DOB: 05/12/42

Id: DCM / Lin: DCM / Id: ID
W: 3498 L: 2000



Case Study Adapted from Addressing HIV Care and Transgender Communities

[redacted], a 33 year old [redacted] male to female transgender person, wants to look into services that you provide. She has not legally changed her name so her documents display her given male name [redacted]. She is new in transition, dresses in high heels and tight skirts. She produces facial hair (which is exposed). She looks to be very nervous, shy and does not look anyone in the eyes. [redacted] has been diagnosed HIV positive for three years and has just begun a relationship with [redacted] who is HIV negative.

Discussion Questions

1. How can the provider establish a sense of trust with the patient?
2. What can the provider do to gain a sense of [redacted] knowledge about HIV, how it is contracted, and how it is treated?
3. What social support, if any, should [redacted] be offered?
4. How can the provider fulfill his/her ethical responsibility to [redacted] and her partner(s)?
5. Discuss other Cultural Competence issues that may impact retention into care and treatment

Acc:
[redacted]
Img Tm: 14:23:53

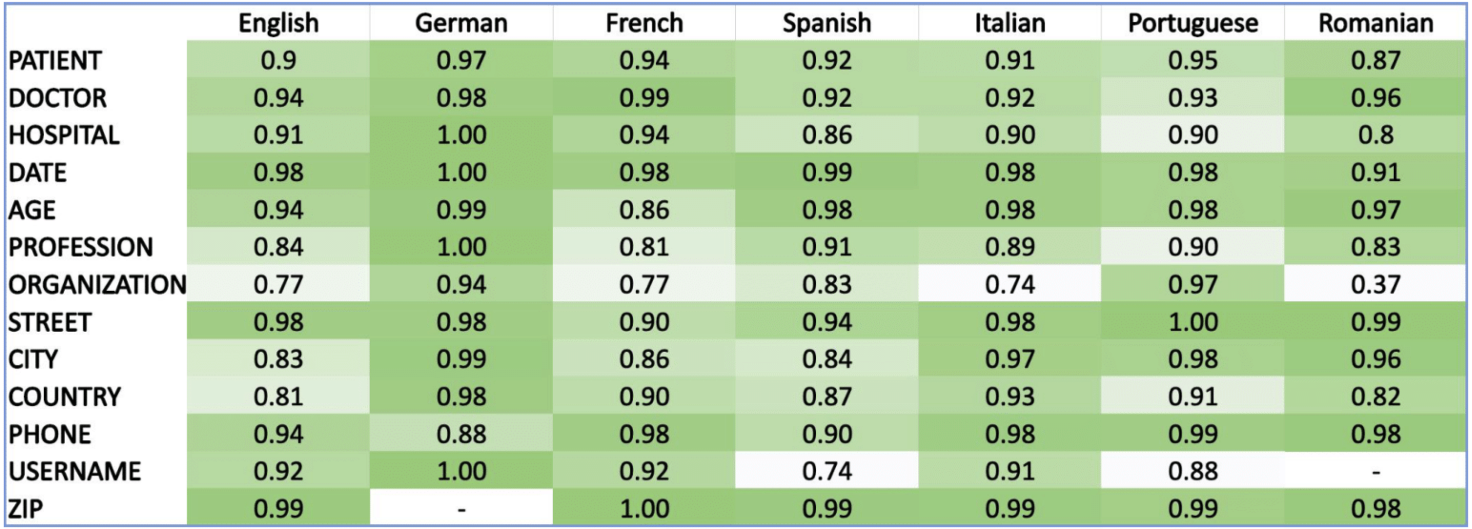
Name: [redacted]
Age: 78
DOB: [redacted]

Id: DCM / Lin: DCM / Id: ID
W: 3498 L: 2000

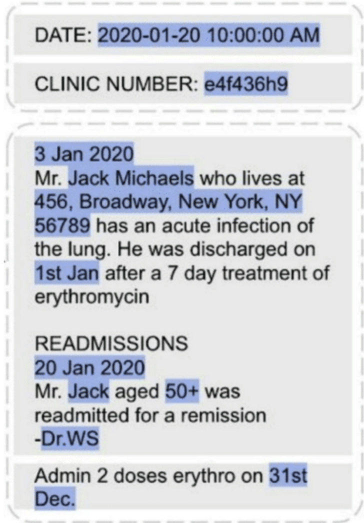
Agenda

- Motivation & background
- John Snow Labs’ de-identification capabilities
- **Technical solution & benchmarks**

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

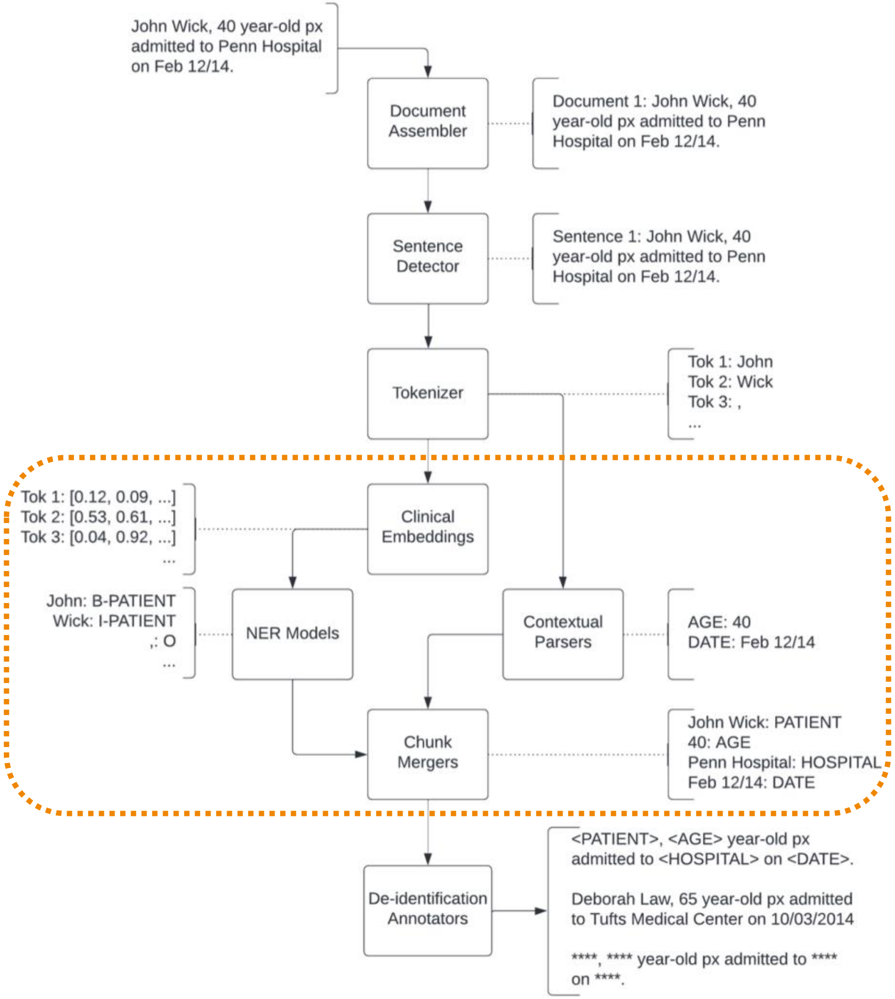
Admin 2 doses erythro on 31st Dec.

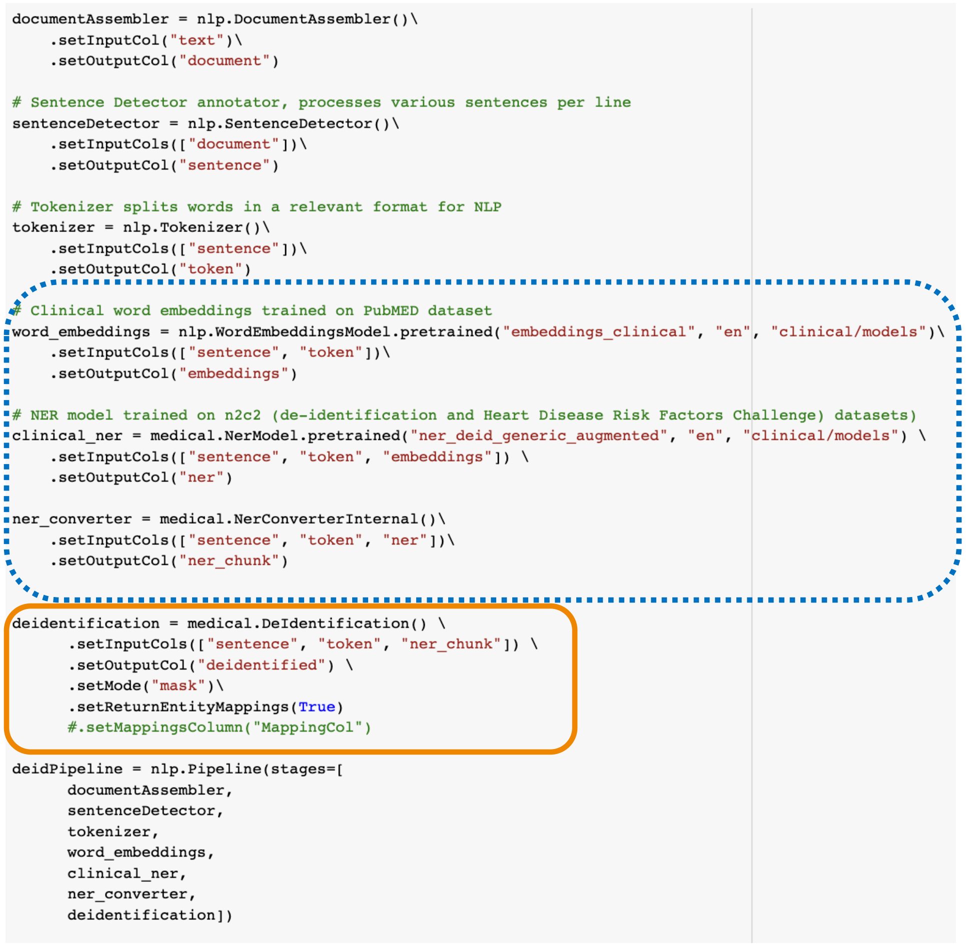
Sub-entities (13-entity)

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)

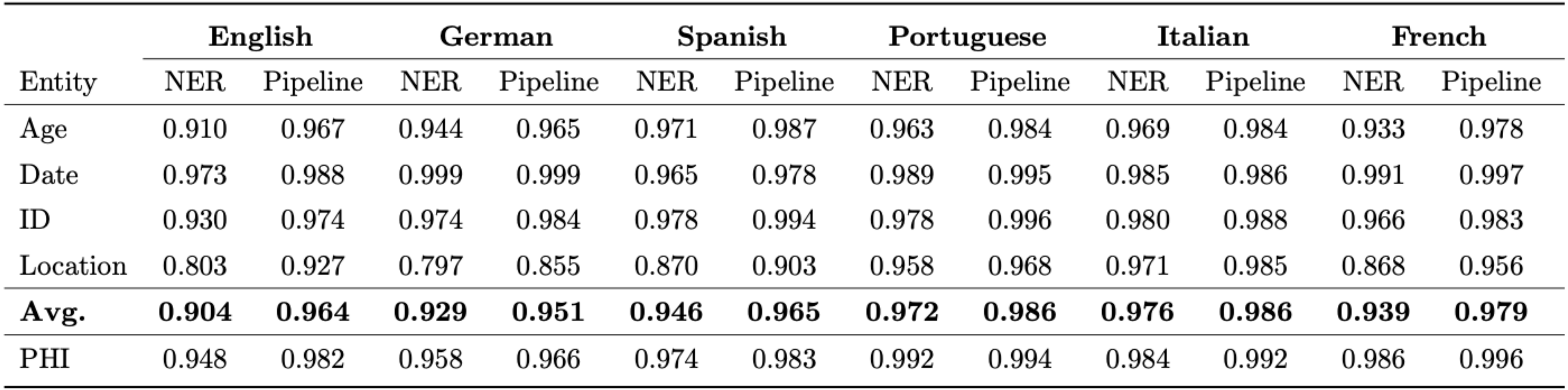
DATE , NAME , LOCATION ,
PROFESSION , CONTACT , AGE , ID


John Wick, 40 year-old px admitted to Penn Hospital on Feb 12/14.
Document Assembler
Document 1: John Wick, 40 year-old px admitted to Penn Hospital on Feb 12/14.
Sentence Detector
Sentence 1: John Wick, 40 year-old px admitted to Penn Hospital on Feb 12/14.
Tokenizer
Tok 1: John
Tok 2: Wick
Tok 3: ,
...
Tok 1: [0.12, 0.09, ...]
Tok 2: [0.53, 0.61, ...]
Tok 3: [0.04, 0.92, ...]
...
Clinical Embeddings
John: B-PATIENT
Wick: I-PATIENT
,: O
...
NER Models
Contextual Parsers
AGE: 40
DATE: Feb 12/14
Chunk Mergers
John Wick: PATIENT
40: AGE
Penn Hospital: HOSPITAL
Feb 12/14: DATE
De-identification Annotators
<PATIENT>, <AGE> year-old px admitted to <HOSPITAL> on <DATE>.
Deborah Law, 65 year-old px admitted to Tufts Medical Center on 10/03/2014
****, **** year-old px admitted to **** on ****.



```
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])
```

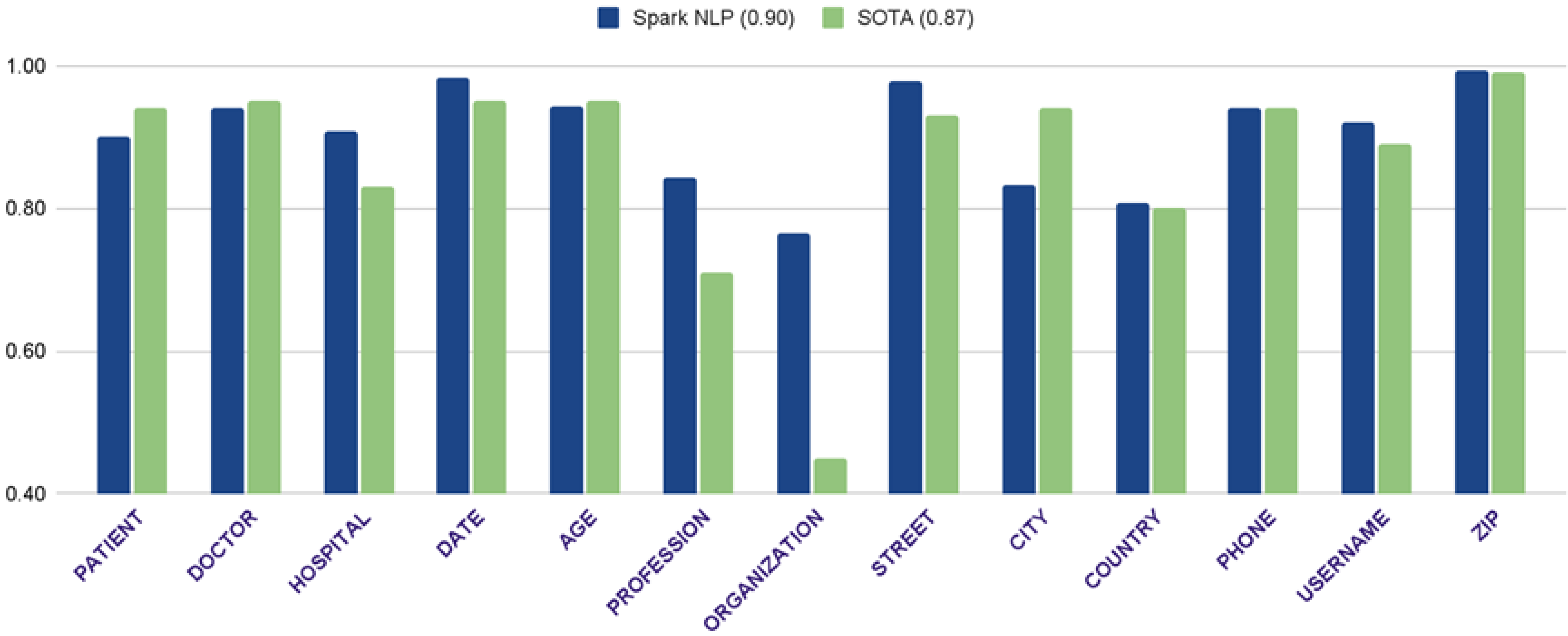
Enriching de-identification pipeline with regex and contextual parser



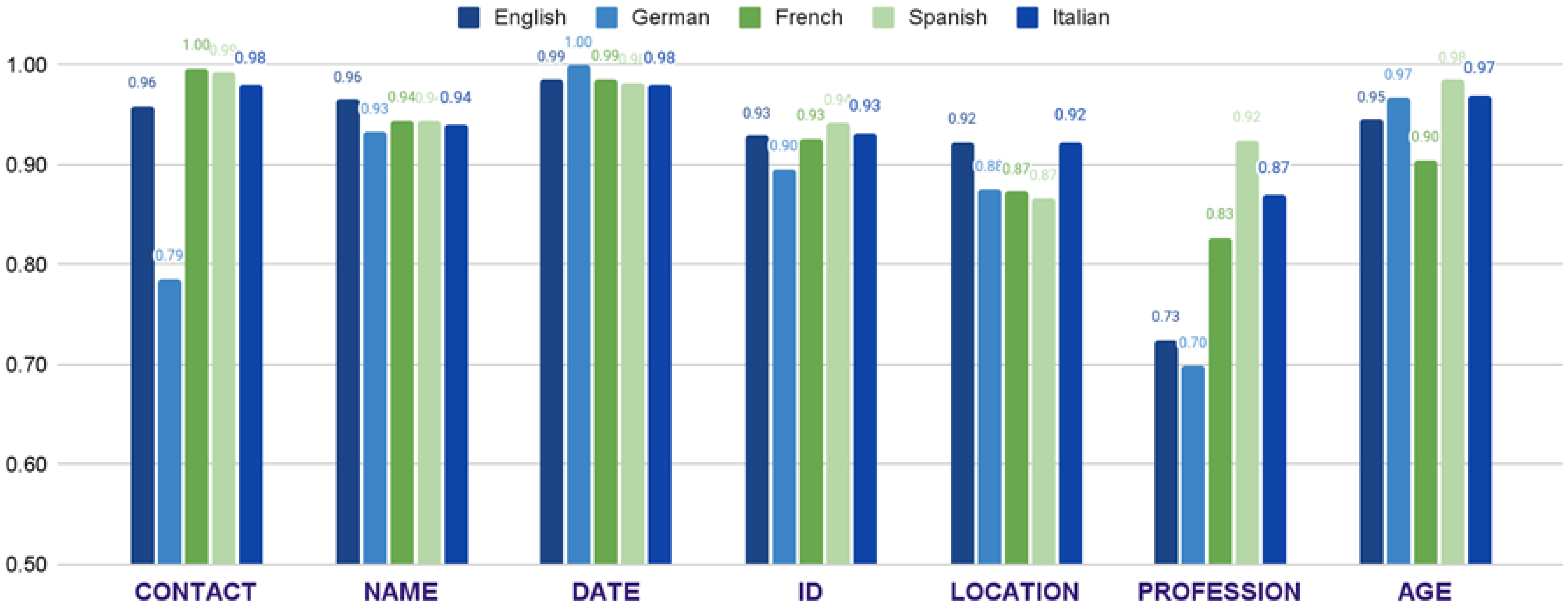
	English		**German**		**Spanish**		**Portuguese**		**Italian**		**French**	
Entity	NER	Pipeline	NER	Pipeline	NER	Pipeline	NER	Pipeline	NER	Pipeline	NER	Pipeline
Age	0.910	0.967	0.944	0.965	0.971	0.987	0.963	0.984	0.969	0.984	0.933	0.978
Date	0.973	0.988	0.999	0.999	0.965	0.978	0.989	0.995	0.985	0.986	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



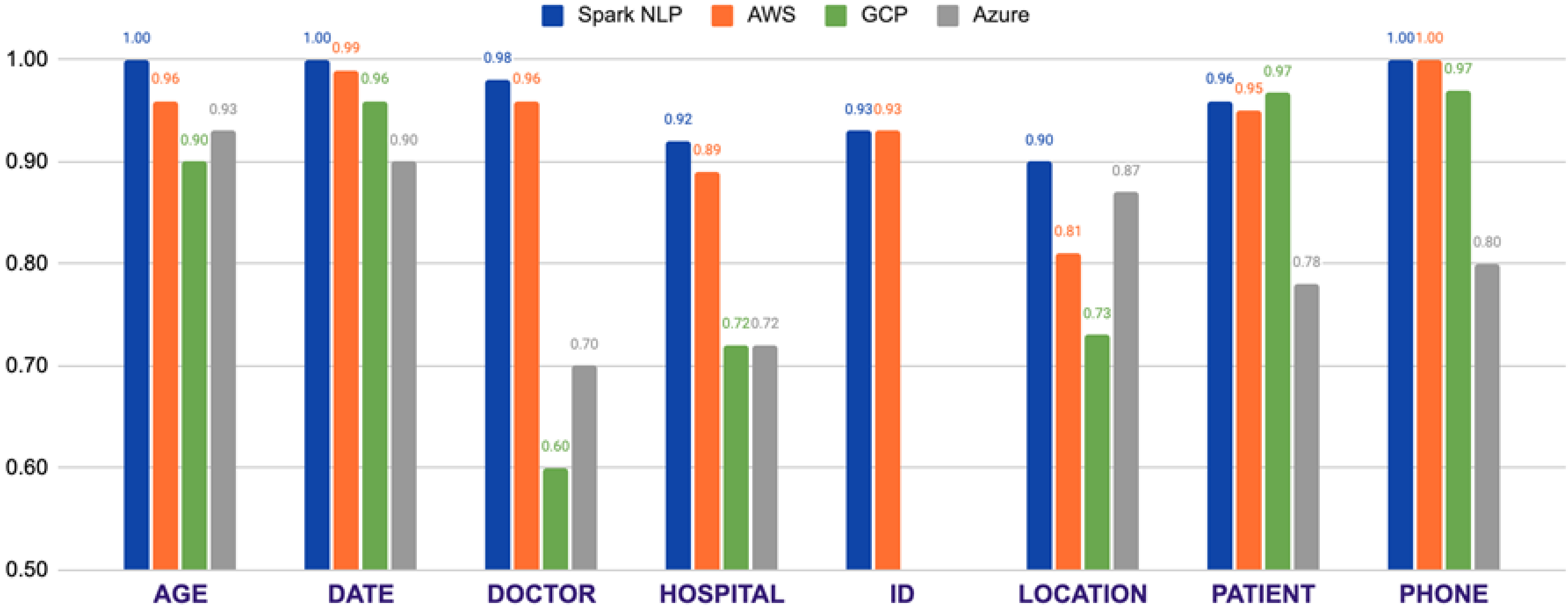
Deidentification Benchmarks



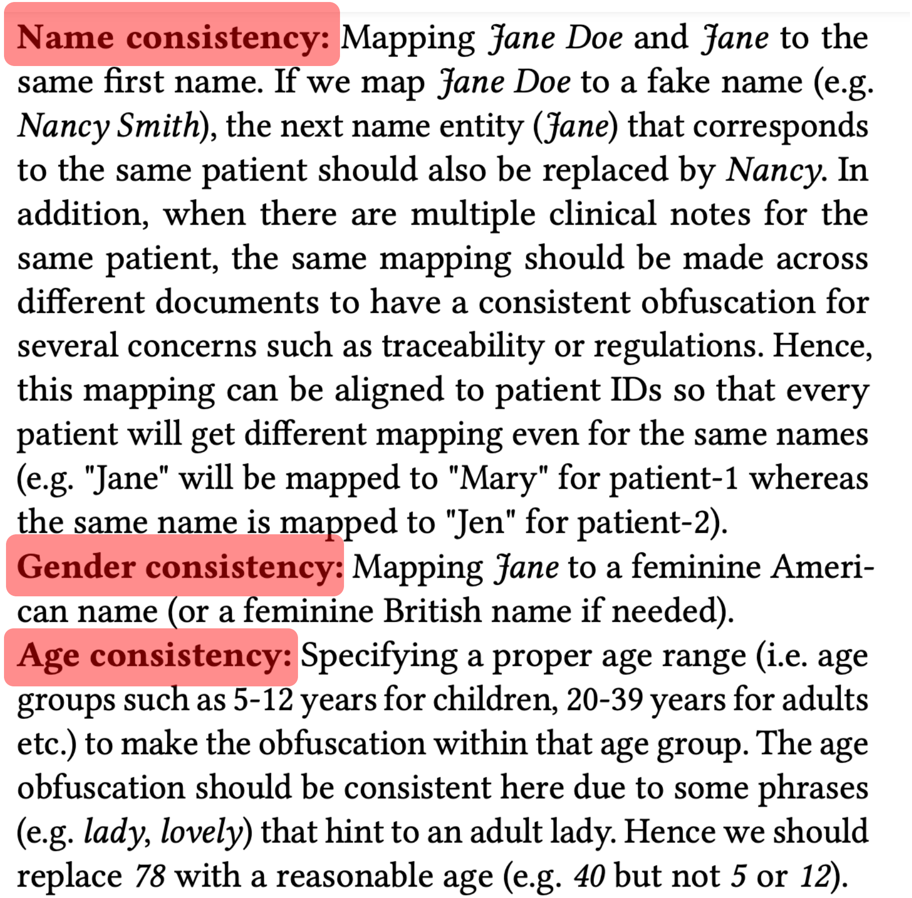
Generic

(7-entity)

De-Identification Benchmarks (en)



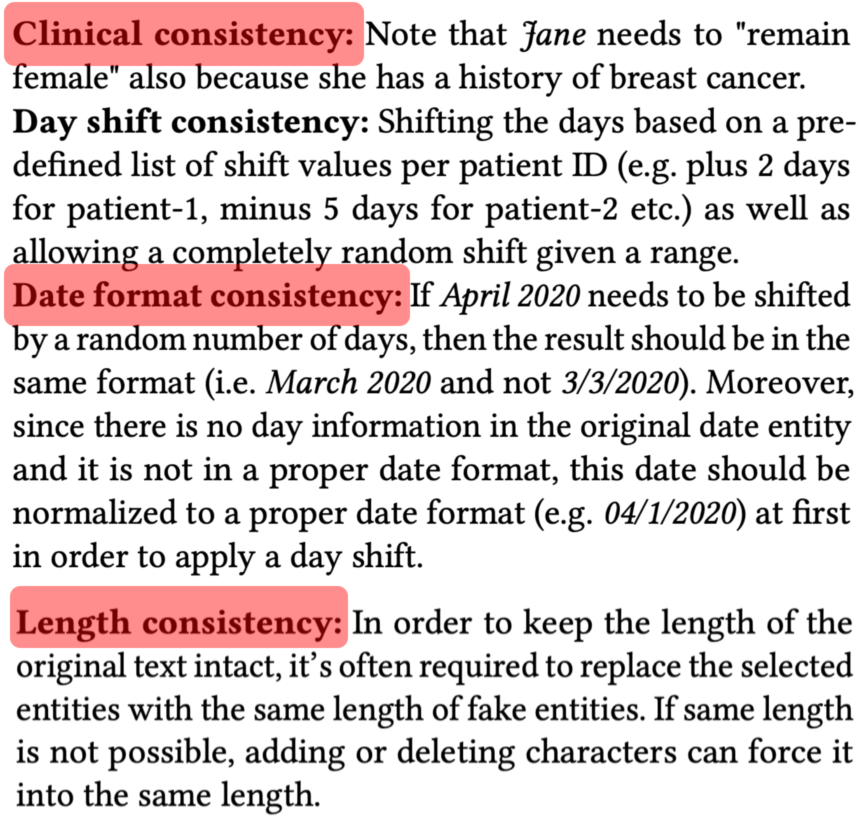
Obfuscation Consistency



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*).



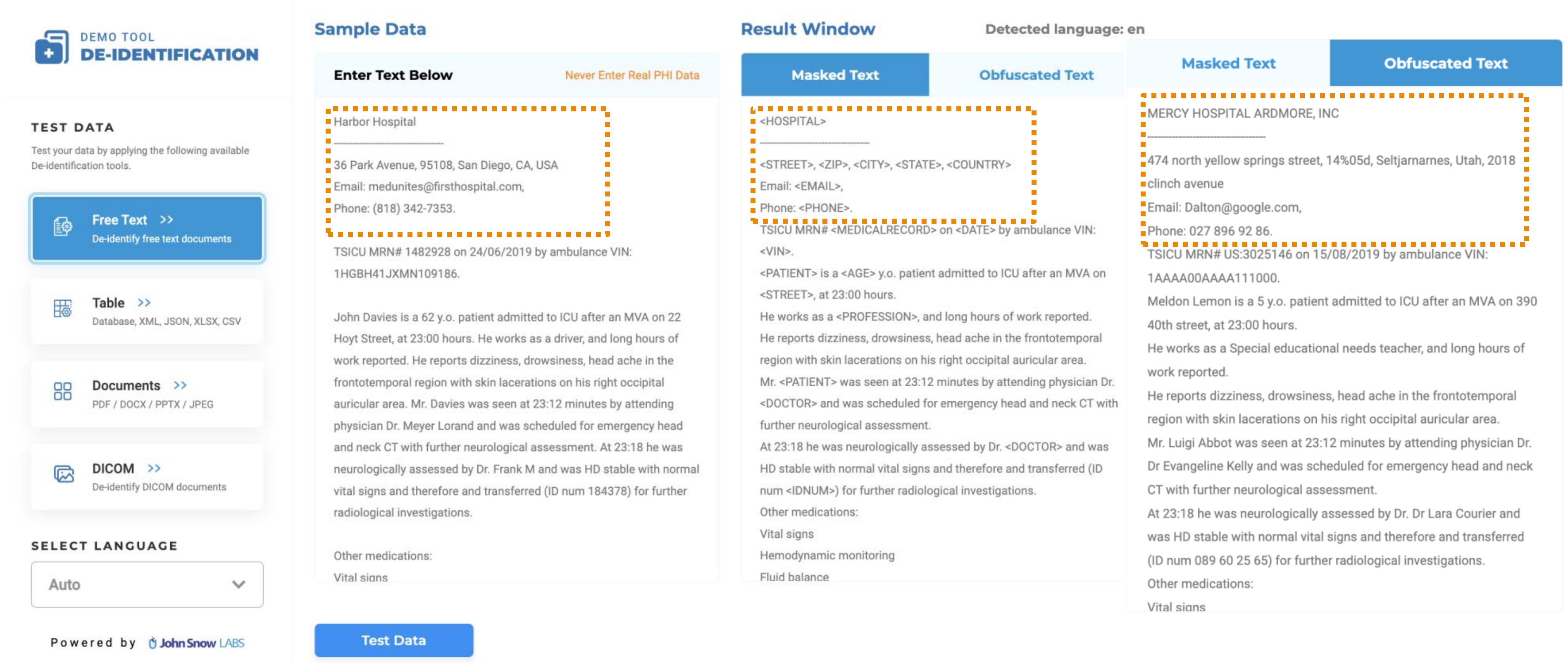
Clinical consistency: Note that *Jane* needs to "remain female" also because she has a history of breast cancer.

Day shift consistency: Shifting the days based on a pre-defined 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.

Live app --> https://www.johnsnowlabs.com/deidentification



Test your data by applying the following available De-identification tools.

Sample Data

Harbor Hospital

------------------------------

36 Park Avenue, 95108, San Diego, CA, USA

Email: medunites@firsthospital.com,

Phone: (818) 342-7353.

TSICU MRN# 1482928 on 24/06/2019 by ambulance VIN: 1HGBH41JXMN109186.

John Davies is a 62 y.o. patient admitted to ICU after an MVA on 22 Hoyt Street, at 23:00 hours. He works as a driver, and long hours of work reported. He reports dizziness, drowsiness, head ache in the frontotemporal region with skin lacerations on his right occipital auricular area. Mr. Davies was seen at 23:12 minutes by attending physician Dr. Meyer Lorand and was scheduled for emergency head and neck CT with further neurological assessment. At 23:18 he was neurologically assessed by Dr. Frank M and was HD stable with normal vital signs and therefore and transferred (ID num 184378) for further radiological investigations.

Other medications:

Vital signs

Masked Text

<HOSPITAL>

------------------------------

<STREET>, <ZIP>, <CITY>, <STATE>, <COUNTRY>

Email: <EMAIL>,

Phone: <PHONE>.

TSICU MRN# <MEDICALRECORD> on <DATE> by ambulance VIN: <VIN>.

<PATIENT> is a <AGE> y.o. patient admitted to ICU after an MVA on <STREET>, at 23:00 hours.

He works as a <PROFESSION>, and long hours of work reported.

He reports dizziness, drowsiness, head ache in the frontotemporal region with skin lacerations on his right occipital auricular area.

Mr. <PATIENT> was seen at 23:12 minutes by attending physician Dr. <DOCTOR> and was scheduled for emergency head and neck CT with further neurological assessment.

At 23:18 he was neurologically assessed by Dr. <DOCTOR> and was HD stable with normal vital signs and therefore and transferred (ID num <IDNUM>) for further radiological investigations.

Other medications:

Vital signs

Hemodynamic monitoring

Fluid balance

Obfuscated Text

MERCY HOSPITAL ARDMORE, INC

------------------------------

474 north yellow springs street, 14%05d, Seltjarnarnes, Utah, 2018 clinch avenue

Email: Dalton@google.com,

Phone: 027 896 92 86.

TSICU MRN# US:3025146 on 15/08/2019 by ambulance VIN: 1AAAA00AAAA111000.

Meldon Lemon is a 5 y.o. patient admitted to ICU after an MVA on 390 40th street, at 23:00 hours.

He works as a Special educational needs teacher, and long hours of work reported.

He reports dizziness, drowsiness, head ache in the frontotemporal region with skin lacerations on his right occipital auricular area.

Mr. Luigi Abbot was seen at 23:12 minutes by attending physician Dr. Dr Evangeline Kelly and was scheduled for emergency head and neck CT with further neurological assessment.

At 23:18 he was neurologically assessed by Dr. Dr Lara Courier and was HD stable with normal vital signs and therefore and transferred (ID num 089 60 25 65) for further radiological investigations.

Other medications:

Vital signs

Presented by John Snow LABS

Thank you.