Agenda

PART 1

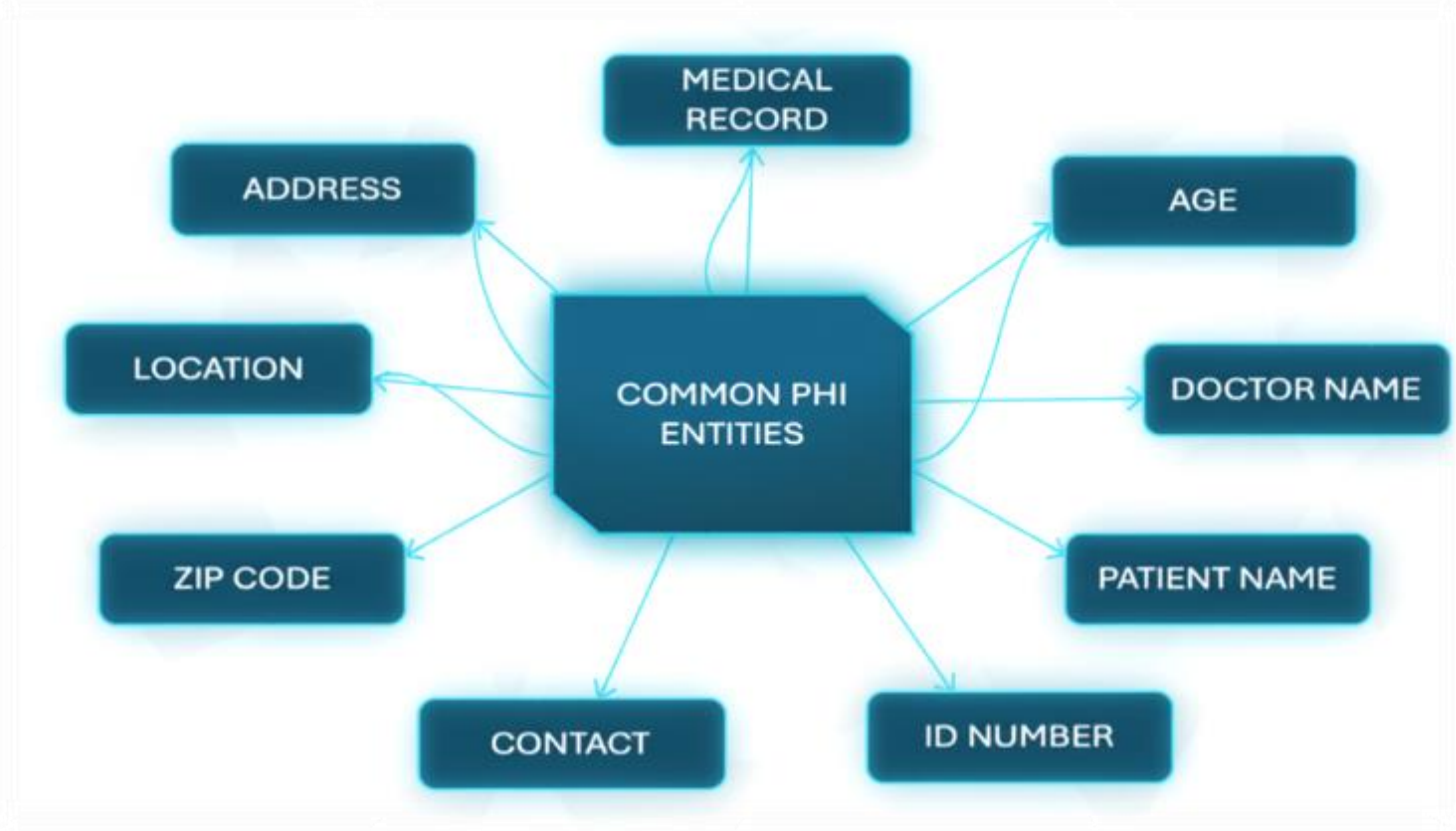
- What is PHI/PII and HIPAA
- HIPAA privacy rule
- What is Deidentified Patient Data
- Why is it Required?
- Peer-Reviewed Papers
- De-Identify Unstructured Clinical Text

PART 2

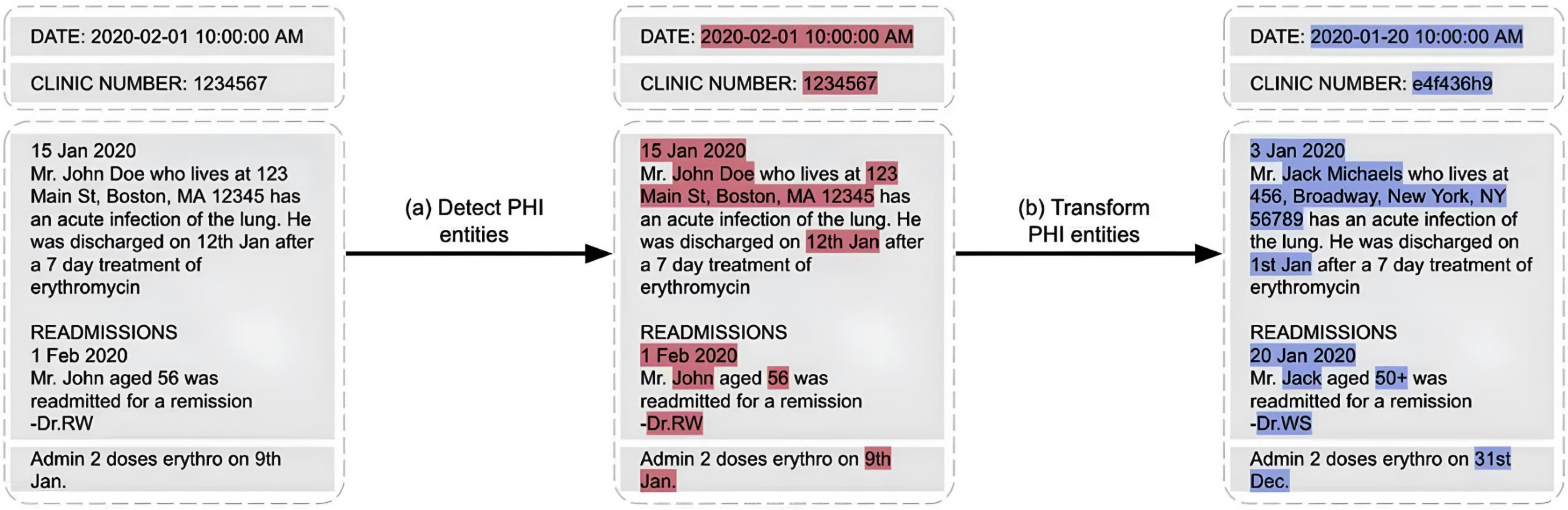
- De-Identify Unstructured Clinical Text
- De-identify structured data

What is PHI/PII and HIPAA

- Personally Identifiable Information (PII) and Protected Health Information (PHI) refer to any data that can be used to identify an individual, such as names, addresses, phone numbers, social security numbers, or medical records
- PHI (Protected Health Information) is defined under HIPAA (Health Insurance Portability and Accountability Act, USA).
- PHI includes all information that can identify an individual and that is related to their health condition, treatment, or payments.



What is Deidentified Patient Data



De-Identification process identifies potential pieces of content with personal information about patients and removes them by replacing them with semantic tags or fake entities.

Why is it Required?

Key Reasons:

- ❑ Privacy Protection
- ❑ Research & Data Sharing
- ❑ Legal Compliance
- ❑ Avoiding Heavy Penalties

Organization	Penalty (USD)	Violation Summary	Timeline	Violated Rules	Source
Anthem, Inc.	💰 **$16,000,000**	Between 2014–2015, Anthem suffered a massive cyberattack that exposed the electronic health information (PHI) of nearly 79 million individuals. This remains the largest HIPAA penalty ever issued.	**Breach:** 2014–2015 **Fine:**October 2018	**Privacy & Security Rules**	HHS.gov – OCR Press Release
Premera Blue Cross	💰 **$6,850,000**	A series of cyberattacks in 2014–2015 compromised the PHI of approximately 10.4 million individuals. The organization agreed to a $6.85 million settlement with the OCR in September 2020.	**Breach:** 2014–2015 **Settlement:**September 2020	**Privacy & Security Rules**	HHS.gov – OCR Press Release

The Data De-identification Process

John Snow LABS

1 Analyze

Human

- Risk analysis
- Legal requirements review

- HIPAA Safe Harbor, HIPAA Expert Determination
- CCPA
- GDPR pseudoanonymization, GDPR anonymization

- Quality assurance strategy & process

2 Identify

Software

- ID, name, email, patient ID, SSN, credit card, address, birthday, phone, URL, license number

- Physician name, hospital name, profession, employer, affiliation

- Racial or ethnic origin, religion, political or union affiliation, biometric or genetic data, sexual practice or orientation

3 Measure

Human

- Cleanroom AI Platform (on-site)
- Annotation tool
- Active learning
- Accuracy Measurement & agreement processes
- Correct sampling
- Multi-lingual

4 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)

5 Monitor

Human

- Ongoing measurement & model improvement
- Missed sensitive data
- Incident response
- GDPR & CCPA requests
- Emergency unblinding
- Audits

Receive raw data

Deliver de-identified data

Peer-Reviewed Papers

Can Zero-Shot Commercial APIs Deliver Regulatory-Grade Clinical Text DeIdentification?

Accepted at Text2Story Workshop at ECIR 2025

"John Snow Labs' Medical Language Models solution achieves the highest accuracy, with a 96% F1-score in protected health information (PHI) detection, **outperforming Azure (91%), AWS (83%), and GPT-4o (79%)**."

Beyond Accuracy: Automated De-Identification of Large Real-World Clinical Text Datasets

Machine Learning for Health (ML4H) 2023

"**The proposed system makes 50%, 475%, and 575% fewer errors than the comparable AWS, Azure, and GCP services respectively while also outperforming ChatGPT by 33%.** It exceeds 98% coverage of sensitive data across 7 European languages, without a need for fine tuning."

Accurate Clinical and Biomedical Named Entity Recognition at Scale

Software Impacts, July 2022

"**Establishes new state-of-the-art accuracy on 7 of 8 well-known benchmarks**, including the 2014 n2c2 de-identification challenge. This implementation outperforms the accuracy of solutions such AWS Medical Comprehend and Google Cloud Healthcare API by a large margin (8.9% and 6.7% respectively)."

Benchmark from "Can Zero-Shot Commercial APIs Deliver Regulatory-Grade Clinical Text De-Identification?

John Snow LABS

Metric / Entity	Healthcare NLP			Azure			Amazon			GPT-4o			Claude 3.7 Sonnet		
	Precision	Recall	F1-score	Precision	Recall	F1-score	Precision	Recall	F1-score	Precision	Recall	F1-score	Precision	Recall	F1-score
AGE	0.96	1.00	**0.98**	0.94	0.45	0.61	1.00	0.41	0.58	0.87	0.50	0.64	0.73	0.55	0.63
CONTACT	0.96	0.97	**0.97**	0.73	0.88	0.80	0.78	0.72	0.75	0.67	0.53	0.59	0.74	0.43	0.54
DATE	0.97	0.99	**0.98**	0.91	0.99	0.95	0.90	0.97	0.93	0.79	0.72	0.75	0.84	0.83	0.83
IDNUM	0.98	0.94	**0.96**	0.78	0.93	0.85	0.95	0.86	0.91	0.70	0.92	0.80	0.70	0.95	0.80
LOCATION	0.93	0.92	**0.93**	0.89	0.87	0.88	0.52	0.74	0.61	0.82	0.72	0.76	0.79	0.87	0.83
NAME	0.92	0.94	**0.93**	0.92	0.89	0.90	0.85	0.76	0.80	0.79	0.82	0.80	0.82	0.86	0.84
O	1.00	1.00	**1.00**	1.00	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99
Macro Avg	0.96	0.97	**0.96**	0.88	0.86	0.85	0.86	0.78	0.80	0.80	0.74	0.76	0.80	0.78	0.78
Non-PHI	1.00	1.00	**1.00**	1.00	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99
PHI	0.96	0.97	**0.96**	0.91	0.92	0.91	0.81	0.85	0.83	0.81	0.77	0.79	0.81	0.84	0.83
Macro Avg	0.98	0.98	**0.98**	0.95	0.96	0.95	0.90	0.92	0.91	0.90	0.88	0.89	0.90	0.92	0.91
cost per 1M doc	**$2,418**			$13,125			$14,525			$21,400			$23,330		

Healthcare NLP, Azure, Amazon GPT-4o and Claude 3.7 Sonnet PHI Recognition and Benchmark Comparison (Sample size: 45172 PHI entities).

De-Identify Unstructured Clinical Text

John Snow LABS

	English	German	French	Spanish	Italian	Portuguese	Romanian	Arabic
PATIENT	0.9	0.97	0.94	0.92	0.91	0.95	0.87	0.87
DOCTOR	0.94	0.98	0.99	0.92	0.92	0.93	0.96	0.93
HOSPITAL	0.91	1.00	0.94	0.86	0.90	0.90	0.80	0.83
DATE	0.98	1.00	0.98	0.99	0.98	0.98	0.91	0.98
AGE	0.94	0.99	0.86	0.98	0.98	0.98	0.97	0.97
PROFESSION	0.84	1.00	0.81	0.91	0.89	0.90	0.83	0.91
ORGANIZATION	0.77	0.94	0.77	0.83	0.74	0.97	0.37	0.82
STREET	0.98	0.98	0.90	0.94	0.98	1.00	0.99	0.98
CITY	0.83	0.99	0.86	0.84	0.97	0.98	0.96	0.97
COUNTRY	0.81	0.98	0.90	0.87	0.93	0.91	0.82	0.94
PHONE	0.94	0.88	0.98	0.90	0.98	0.99	0.98	0.92
USERNAME	0.92	1.00	0.92	0.74	0.91	0.88	-	1.00
ZIP	0.99	-	1.00	0.99	0.99	0.99	0.98	0.97

De-identify structured data

Original Table

ID	SSN	NAME	DOB	POB	ADDRESS
1002301	547-46-9390	Alex Williams	1970-01-01	Salt Lake City, Utah	615 Walton Street, Salt Lake City, UT, 84111
4052191	433-10-5021	David Smith	1955-10-13	Charleston, SC	3261 Broadway Street, Charleston, SC, 29424
7017021	322-21-1197	Mary Johnson	1965-04-03	Fayetteville, New York	4116 Confederate Drive, Fayetteville, NY, 13066

Masked Table

ID	SSN	NAME	DOB	POB	ADDRESS
<MEDICALRECORD>	<SSN>	<PATIENT>	<DATE>	<CITY>, <STATE>	<STREET>, <CITY>, <STATE>, <ZIP>
<MEDICALRECORD>	<SSN>	<PATIENT>	<DATE>	<CITY>, <STATE>	<STREET>, <CITY>, <STATE>, <ZIP>
<MEDICALRECORD>	<SSN>	<PATIENT>	<DATE>	<CITY>, <STATE>	<STREET>, <CITY>, <STATE>, <ZIP>

Obfuscated Table

ID	SSN	NAME	DOB	POB	ADDRESS
T8608038	049-65-6118	Dorcas Clara	1970-02-10	Savageville, Ohio	Amsinckstrasse 9, Savageville, Virginia, 00008
Y293561	654-56-5824	Secundino Blades	1955-11-22	Jacksonhaven, New Mexico	Jamesland, Jacksonhaven, New Mexico, 00006
S1330746	063-21-4576	Lorita Medal	1965-05-08	Opole, California	4855 Blue Diamond Road, Opole, Iowa, 00002