phuse·share
connect·share·advance

Single Day Event

Unleashing the Potential of NLP in Clinical Data Analysis and Standardization

Meckee Munka, Rahul Pandya
ICON Strategic Solutions

The components of this presentation

- Clinical Trials : Changing times
- The 20-20 of structure and standards
- Structure the unstructured
- Learn from the past
- Predict the future
- Technology joins hands with standards
- NLP from FDA’s perspective
- Case Study: SDTM Mapping based on NLP and ML
- Challenges & Possible Solution
- Looking forward to

Clinical Trails: Changing times

Complex clinical trials

Increasing amounts of data

Patient-Centric Technology

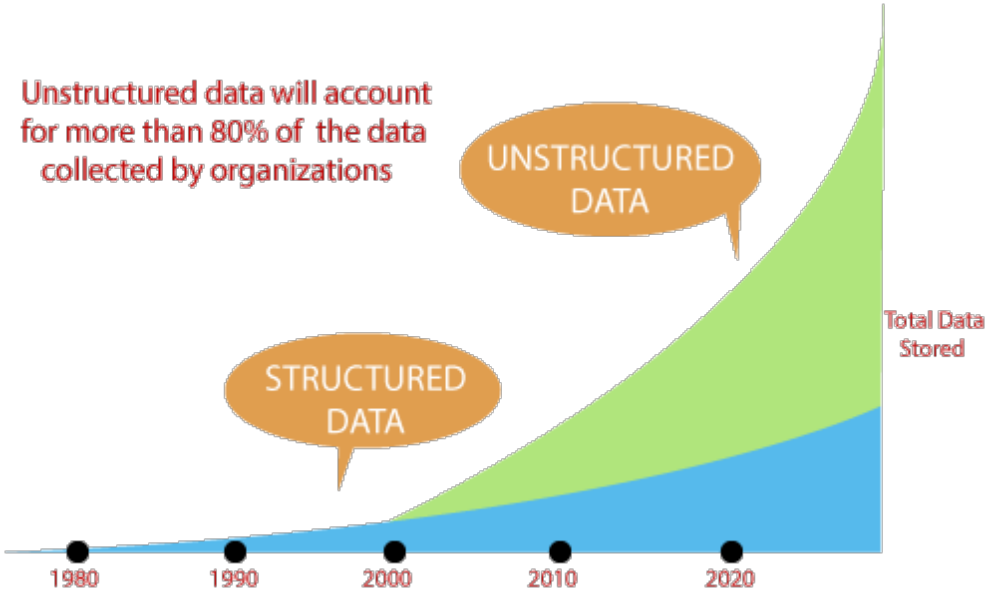
Wearable Technology

The 20-20 of Standards and Structure

The unstructured data are :

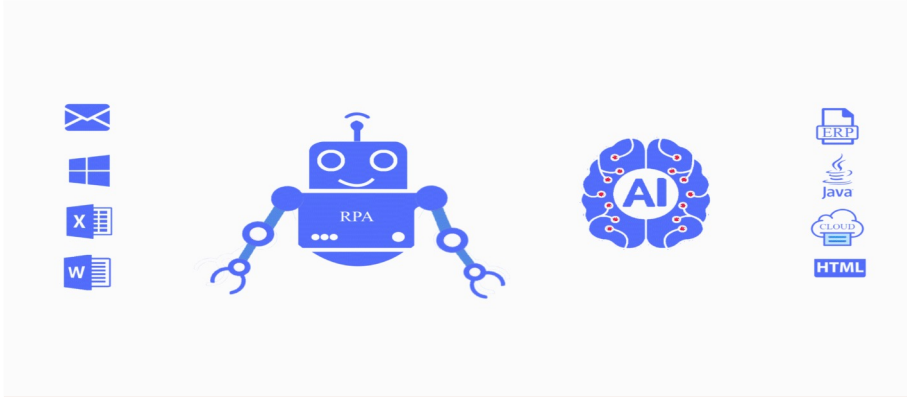
- ✓ typed and written text,
- ✓ radiological images,
- ✓ pathology slides,
- ✓ video,
- ✓ audio, and
- ✓ streaming device data.

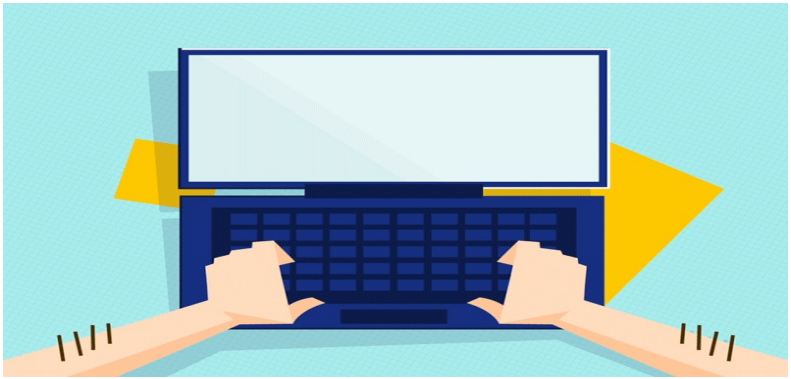
Unstructured data is valuable, but it's difficult to store, search, analyze and share with collaborative teams, especially if they work in different healthcare organizations.



Structure the Unstructured

To help manage this avalanche of data, more pharmaceutical companies are turning to natural language processing (NLP) technology to mine unstructured, text-based documents and convert the data into structured information that can be analyzed by a computer

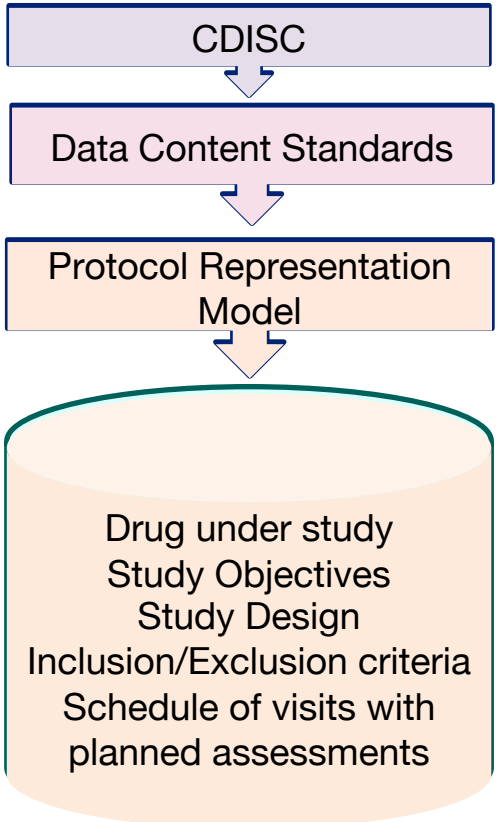




Numerous companies are tapping into NLP technology to explore the potential of unstructured data into clinical analysis, one such instance is of a leading pharmaceutical company where the company used NLP to digitalize historical protocols.

Learn from the past

Pharmaceutical companies have 20 plus years of study protocols.



Till now this has been largely ignored because they were locked away across disparate systems in mixed format documents that would be far too resource intensive to parse for every new study.

Now imagine all this data from clinical protocols into normalized structured data suitable for analysis or machine learning algorithms.

Predict the future

Will it not accelerate answers to questions like

What study designs were most efficient in breast cancer therapy?

What inclusion/exclusion criteria helped to meet objectives in a 75+ population?

What sites failed to recruit in patients with arthritis

The potential is huge.

From clinical data standard's perspective, on the very surface this integration of protocols has clubbed two of the CDISC's Data content standards

Protocol Representation Model
Study Design Model

Technology joins hands with standards

Healthcare Natural Language APIs

Unstructured medical records and
Real World Data from
FHIR resource

Medical Vocabularies	Medical categories	Functional categories	Relationship between entities
Example: LOINC, MeSH, SNOMED, RxNORM	Example: BM_UNIT – Body Measurement Unit MED_DOSE- Medication Dose	Example: CLINCAL_HISTORY, FAMILY_HISTORY,FAMILY_MEMBER, LIKELY, SOMEWHAT LIKELY	Example: LABORATORY_DATA LAB_RESULT, LAB_UNIT, LAB_VALUE
		Investigator	ADaMs

NLP from FDA's perspective

FDA recognizes the use of AI/ML in drug development lifecycle including:

- Clinical trial designs,
- DHTs (Digital Health Technologies) and
- RWD analytics.

FDA supports AI/ML in submission

Submission have grown using AI/ML references

FDA also plans to use claims and EHR data for its 5 years Sentinel strategic plan – which will explore AI/ML along with NLP to automate extraction of unstructured clinical notes for computable phenotyping

FDA is exploring NLP and ML to classify free-text narratives in FDA Adverse Events Reporting Systems into structured medical dictionary.

Case Study: SDTM Mapping based on NLP and ML

Objective
To explore approach of NLP and ML in SDTM Mapping for clinical data standardisation

Method
Adverse Event domain was selected as proof of concept of NLP and ML based SDTM mapping

Data Prepared from
Raw Datasets from 20 studies of 3 diff CRO, 2 EDC systems, Phase1 to Phase3
Metadata : SDTM-IG 3.2, CDASH 2.0

Generate NLP models that represent the mapping specification of raw variable to target SDTM and SUPPQUAL variables:

MODEL A

Built based on tokenized terms using the traditional statistical method

MODEL B

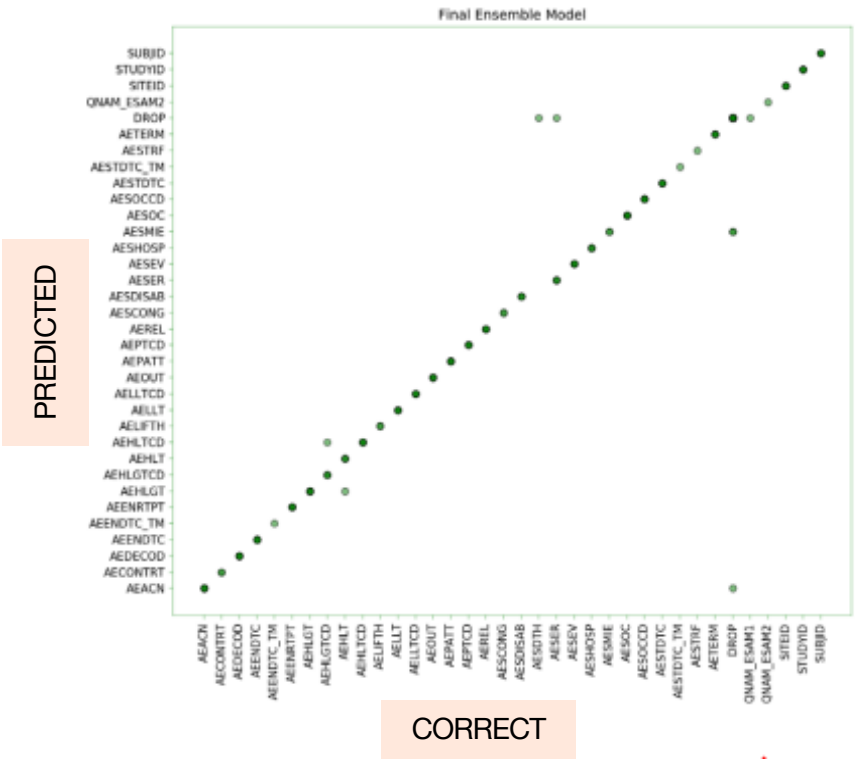
Pre trained word embedding based on feedforward neural natural language.

MODEL C

ensemble of Model A and Model B

Model A, Model B and Model C were evaluated with logistic regression for model performance. The language **Model C** has highest accuracy of **82 %** while Model A had lowest accuracy of 64%

Case Study: SDTM Mapping based on NLP and ML



Result

Ensembled model (Model C) of NLP with ML shows prominent results, where its achieved accuracy of 97.5%

Conclusion:

The study demonstrated that machine learning based mapping using Natural Language Processing (NLP) can be used to assist in automated SDTM mapping.

The model can be further improved by using a larger word embedding training on CDISC document, the choice of different NLP method with a rule-based algorithm.

Challenges & Possible Solution

Data Quality & Ambiguity

Clinical data is often unstructured and contains medical jargon, acronyms, and abbreviations, which can make it difficult to process and analyse.

Limited availability of Annotated data

NLP models require annotated data to train and improve. However, in the healthcare domain, annotated data is often limited due to privacy concerns.

Difficult to Interpret

In the healthcare domain, explainability is crucial for ensuring that decisions made by NLP models are clinically relevant and accurate.

Ethical & Legal

The use of NLP in clinical data standardization and solutions raises ethical and legal considerations, particularly about patient privacy, data protection, and informed consent.

Challenges & Possible Solution

Solution: Build **custom NLP healthcare model** which cater the need of clinical industry more effectively and accurately.

This model is something which is developed with focus on,

- Trained on domain-specific data including the standard regulatory guidelines
- Use data encryption and secure storage practices to protect user data. Built annotation with data encryption which helps NLP model to get trained.
- Use data cleaning and pre-processing techniques to remove noise and inconsistencies from the data before training the model.

Looking forward to

Thank You

The Global Healthcare Data Science Community

Contact Channels
UK +44 1843 609600
office@phuse.global
phuse.global

Social Media
@phusetwitta
/phusebook
/phusetube
/company/phuse