NLP for Healthcare: Challenges With Processing and De-Identifying Clinical Notes

Chloe Pou-Prom, Vaakesan Sundrelingam

Agenda

- Text data in healthcare
- Sharing data in healthcare
- Why anonymization is important
- What are you looking for in a de-identification tool
- `pydeid`
- Demo

About the speakers

- **Unity Health Toronto** is a healthcare network consisting of 3 hospitals in the Greater Toronto Area.

About the speakers - Chloe

- Chloe works with the **Data Science and Advanced Analytics (DSAA)** team.
- DSAA suits the needs of the hospital, our collaborators, and our partners to *make better decisions*, *increase hospital efficiency*, and *improve patient care and patient outcomes*.
- DSAA works with *clinicians* and *administrative decision-makers* to develop and deploy solutions.

DSAA

- Currently more than 40 active solutions at Unity Health:
  - Predicting patient outcomes for enhanced clinical management
  - Planning for hospital bed capacity
  - Medical imaging AI tools
  - Assignment/scheduling

About the speakers - Vaakesan

- Vaakesan works with the **GEMINI** at Unity Health Toronto.
- GEMINI is a unique big data collaborative supporting cutting-edge quality improvement and research projects.
- The GEMINI study collects, formats, standardizes and analyzes clinical data from hospitals with the aim of improving how healthcare is delivered.

GEMINI

- Data collected from the GEMINI study includes:
  - 1,600,000+ patient admissions
  - 30+ Canadian hospitals
  - 3.8 billion+ data points
- Administrative and clinical data
  - Labs
  - Vitals
  - Imaging
  - Pharmacy

Challenges With Processing and De-Identifying Clinical Notes

Text data in healthcare

- Admission notes
- Radiology reports
- Consult notes
- Nurse notes
- Discharge notes

- Operative summary

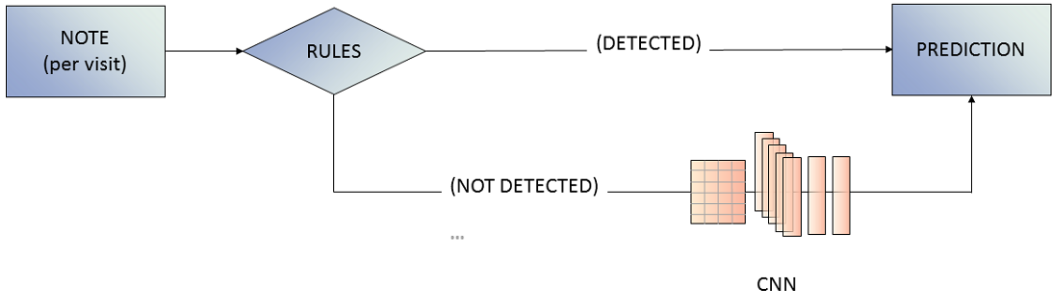
```
GAMGEE, SAMWISE
MRN: 123-4567
D.O.B. Jul-09-1983
DATE OF OR: June 11, 2019
PROCEDURE START TIME: 8:27 p.m.
SURGEON: Dr. Galadriel
ASSISTANTS: Dr. B. Baggins, second clinical fellow; Dr. F. Baggins, PGY5
ANESTHESIA: General.
CLINICAL NOTE: Mr. Gamgee is a 39-year-old male with past medical history
significant for transient ischemic attack and cerebrovascular accident
and type 2 diabetes mellitus. His medication includes Aspirin and Plavix.
She had a fall on June 8, 2019, and he was transferred to a local
hospital and his imaging demonstrated no evidence of intracranial hemorrhag
OPERATIVE NOTE: The patient was brought to the operating room. Briefing
was done. Preoperative antibiotics were given. At the end of the
procedure, all the counts were correct. There were no intraoperative
complications. The patient is being transferred to the intensive care
unit intubated.
```

What can we do with clinical notes?

- Extracting EDSS from clinical notes of patients with Multiple Sclerosis
  - The Multiple Sclerosis (MS) Clinic at St. Michael's Hospital is one of Canada's largest MS clinics.
  - After each visit to the clinic, the MS clinician will dictate a consult note summarizing the visit.
  - The MS Clinic wants to build a research database to monitor trends in disease progression and response to treatments.

What can we do with clinical notes?

- The Expanded Disability Status Scale (EDSS) is a score ranging from 0 to 10 that is commonly-used to quantify and monitor changes in MS-related disability over time.



Read more: "Assessment of Natural Language Processing Methods for Ascertaining the Expanded Disability Status Scale Score From the Electronic Health Records of Patients With Multiple Sclerosis: Algorithm Development and Validation Study" (2022).

What can we do with radiology reports?

- GEMINI has partnered with UofT and the Vector Institute to develop a tool to identify rates of delirium
- Delirium is acute confusion that:
  - affects up to 40 percent of older adults hospitalized for other illnesses
  - Could be preventable in 20 to 40 per cent of cases
- Provide hospitals information about the rates of delirium in the different hospital units, and deploy systems and services for prevention where most needed
- Using radiology reports in a supervised learning task to predict rates of delirium

Predicting delirium using radiology reports

Text data can be difficult to work with

- Flexible formatting
  - e.g., “B Baggins: height - (in) 42 BSA (m2): 2.39 m2 BP (mm Hg): 92/52 HR (bpm): 120”
- Atypical grammar
  - e.g., “imaging showed no evidence of [an] orc bite”

Text data can be difficult to work with

- Language specific to medical domain
  - e.g., “pt has T1 diabetes” vs “T1 vs T2 MRI”
- Misspellings
  - e.g., “repiratoy failure”

Text data can be difficult to work with

- Real-time availability

Procedures database:

procedure_id	start_ts	end_ts
123456	2022-11-22 12:50	2022-11-22 14:50

Notes database:

procedure_id	note_id	ts	note
123456	67890	2022-11-28…	…..

Text data can be difficult to work with

- Real-time availability

```
DATE OF OR: November 24, 2022
PREOPERATIVE DIAGNOSIS: Cholecystitis
POSTOPERATIVE DIAGNOSIS: Cholecystitis
ANESTHESIA: General - given by Dr. Boromir.
SURGEON: Dr. Faramir
ASSISTANTS: Dr. Denethor II, Dr. Theoden
CLINICAL NOTE: Mr. Gimli is a 139-year-old gentleman who presented to the E
with abdominal pain. Investigations were concerning for cholecystitis with
choledocholithiasis.
```

Text data can be difficult to work with

- Different data entry strategies
  - Back-end dictation
  - Front-end dictation
  - Electronic data entry
  - Hand written

Text data can be difficult to work with

- Different data storage strategies
  - Multiple notes per encounter
  - Appending or updating notes
    - Different providers
    - Appending daily updates
    - Follow-up items

Structured

```
Record date: 1067-04-14

Admission Note

PATIENT:  Sméagol/Gollum
MRN:  6827938
ADMIT DATE: 3/11/1067
PCP:  Gandalf the White, MD
ATTENDING PHYSICIAN:  Haldir of Lórien, MD

CHIEF COMPLAINT
Hyponatremia

HISTORY OF THE PRESENT ILLNESS
Sméagol/Gollum is a 143 yo hobbit with a history of schizophrenia, DM2, HTN
presenting with confusion and hyponatremia.

MEDICATIONS ON ADMISSION
```

Unstructured

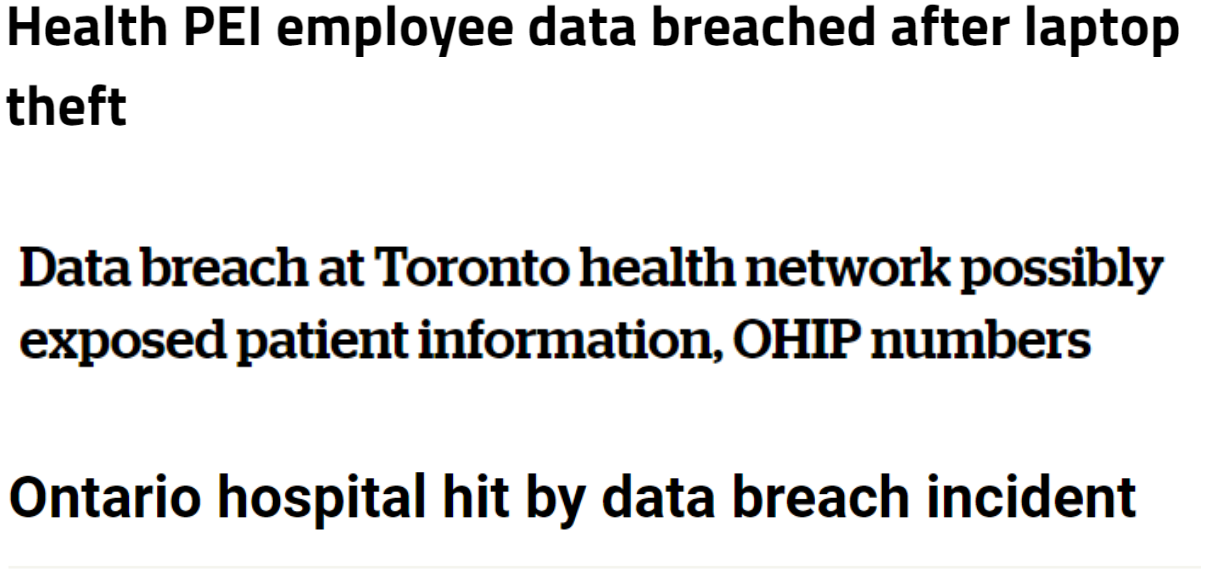
```
Asked by Dr. Goldberry to consult on glycemic management for Mr. Bombadil w
no history of DM but was noted to be hyperglycemic postop.

Mr. Bombadil is post-operative day 1. His family history is positive for di
(mother diagnosed in her 80s). He has required a regular insulin gtt protoc
intermittently postop. He is now on a full liquid diet.

Thank you, please call with any questions, Diabetes Center beeper # 23452.
```

Sharing data in healthcare is difficult

- Getting access is difficult, especially if you're not already working at the hospital or an affiliate research group.



Sharing data in healthcare is *important*

> After reviewing the literature on potential reidentifications of patients in publicly available datasets, we argue that the cost—measured in terms of access to future medical innovations and clinical software—of slowing ML progress is too great to limit sharing data through large publicly available databases for concerns of imperfect data anonymization.

"Global healthcare fairness: We should be sharing more, not less, data", 2022.

Sharing data in healthcare is *important*

> …**personal health information is essential for public health surveillance and health-related research**. The availability of information for such purposes results in enormous benefits for individuals and society at large by improving health-care programs and services and by improving the effectiveness of the health-care system.

"Dispelling the Myths Surrounding De-identification"

Using healthcare data to improve healthcare

- GEMINI (for example) makes data available to healthcare researchers.
- Insights from this research is used to help physicians, health care teams and hospitals gain insights into patient care and improve patient outcomes.

If you're sharing healthcare data, you need to de-identify it!

Why de-identify data?

> De-identification is an essential mechanism for protecting privacy…

"Dispelling the Myths Surrounding De-identification"

- Personal Health Information Privacy Act (PHIPA)
- Research Ethics Board (REB)
- Data Sharing Agreements
- Privacy

Why de-identify data?

- PHIPA:
  - PHIPA permits disclosures of PHI without consent for research purposes if the researcher:
    - prepares a research plan (that meets certain requirements) and
    - a research ethics board (that meets certain requirements) approves the plan

Why de-identify data?

- Research Ethics Board
  - REB provides approval
  - TCPS 2 (Tri-Council Policy Statement on Ethical Conduct for Research Involving Humans)
    - Use fully anonymized data
    - Use data de-identified with a key held by a trusted third party
    - Collect identifiable data, and take measures to de-identify the data as soon as possible.

Why de-identify data?

- Identifiers
  - Direct identifiers: name, address, etc.
  - Indirect identifiers: gender, marital status, location, etc.
  - Rare cases:
    - Extreme ages, rare diagnosis/presentation/adverse event, etc.
    - Cell size of five rule

Why de-identify data?

- Different release models
  - Public release: anyone can download and use the data without any conditions
  - Non-public release: download is limited
  - Semi-public release: combination of public and non-public

Why de-identify data?

- Data sharing agreements
  - help mitigate the risk of re-identification for non-public releases
  - identify the parameters which govern the collection, transmission, storage, security, analysis, re-use, archiving, and destruction of data

Why de-identify data?

- Governance process:
  - regular re-identification risk assessments
  - auditing
  - overlapping data sets
  - response to re-identification attack
  - training
  - …and more!

What do *you* need in a de-identification tool?

What do *you* need in a de-identification tool?

- Speed
  - regex-based > ML-based

	Total time (s)	**chars/s**
Regex	4.84	1,395.59
BiLSTM-CRF (ML)	75.30	23,975.01

What do *you* need in a de-identification tool?

- Accuracy
  - Precision: $\frac{TP}{TP+FP}$
  - Recall: $\frac{TP}{TP+FN}$
  - F1: $2 \cdot \frac{precision \cdot recall}{precision+recall}$
  - Typically: ML-based > regex-based

What do *you* need in a de-identification tool?

	Precision	Recall	F1
Physionet DeID (regex)	0.895	0.698	0.785
PHIlter (regex)	0.786	0.999	0.879
deidentify (ML)	0.959	0.869	0.912

- "Protected Health Information filter (Philter): accurately and securely de-identifying free-text clinical notes" (Norgeot et al., 2020)
- "Comparing Rule-Based, Feature-Based and Deep Neural Methods for De-Identification of Dutch Medical Records." (Trienes et al., 2020)

What do *you* need in a de-identification tool?

- Interpretability
  - The "random replacement" paradigm achieves perfect recall
  - "Using word embeddings to improve the privacy of clinical notes" (Abdalla, 2020)

What do *you* need in a de-identification tool?

- Context

- Healthcare vs. non-healthcare

```
Elrond, Lord
DATE OF OR: 13-September-2008
PREOPERATIVE DIAGNOSIS: Chronic liver abscess.
POSTOPERATIVE DIAGNOSIS: Chronic liver abscess.
OHIP NUMBER: 1234-567-000-AR
...
The gall bladder was irrigated. And then two
Jackson-Pratt drains were laid into the right
upper quadrant and sutured into position.
...
```

What do *you* need in a de-identification tool?

- Context
- US vs. Canada

```
Elrond, Lord
DATE OF OR: 13-September-2008
PREOPERATIVE DIAGNOSIS: Chronic liver abscess.
POSTOPERATIVE DIAGNOSIS: Chronic liver abscess.
OHIP NUMBER: 1234-567-000-AR
...
The gall bladder was irrigated. And then two
Jackson-Pratt drains were laid into the right
upper quadrant and sutured into position.
...
Electronically signed by Bilbo Baggins MD,
The Shire General Hospital, 11103
```

What do *you* need in a de-identification tool?

- Context
- US vs. Canada

```
Elrond, Lord
DATE OF OR: 13-September-2008
PREOPERATIVE DIAGNOSIS: Chronic liver abscess.
POSTOPERATIVE DIAGNOSIS: Chronic liver abscess.
OHIP NUMBER: 1234-567-000-AR
...
The gall bladder was irrigated. And then two
Jackson-Pratt drains were laid into the right
upper quadrant and sutured into position.
...
Electronically signed by Eowyn of Rohan MD,
Rivendell Hospital, M5B 1T8
```

What do *you* need in a de-identification tool?

- Cost
  - Commercial vs open-source solutions
  - Cost structure

What do *we* need in a de-identification tool?

- Speed:
  - Data Sharing Agreements
- Accuracy:
  - REB Approval
- Interpretability:
  - Generalizable for research
  - Readable for annotation

What do *we* need in a de-identification tool?

- Context:
  - Canadian Healthcare
- Cost:
  - High volume of data
  - Open source

pydeid

About `pydeid`

- `pydeid` is a Python-based de-identification software that identifies and replaces personal health information (PHI) in free-text clinical data.
- Our use of the program is vetted through Privacy and the Research Ethics Board (REB)
- **Installation**: coming soon!
  - Follow github.com/GEMINI-Medicine

Motivation and background

- Regex-based (for speed), open source tool (for cost)
  - Philter
  - Physionet Deid
- Modified the Physionet Deid tool for a Canadian healthcare context
  - Received REB approval
- Ran into scalability limitations
- Difficult to manage perl codebase
- Refactored into a `python` package

pyDeid

- De-identification

```
>> from pyDeid import deid_string
>> original_string = 'Arwen Undómiel was born in Rivendell on March 1st, 24
>> phi, new_string = deid_string(original_string)

>> print(new_string)
'Jane Doe was born in London on 1976/6/3'
```

pyDeid

- De-identification

```
>> from pyDeid import deid_string
>> original_string = 'Arwen Undómiel was born in Rivendell on March 1st, 24
>> phi, new_string = deid_string(original_string)

>> print(new_string)
'Jane Doe was born in London on 1976/6/3'

>> print(phi)
 [{'phi_start': 0,
  'phi_end': 4,
  'phi': 'Arwen',
  'surrogate_start': 0,
  'surrogate_end': 3,
  'surrogate': 'Jane',
  'types': ['Female First Name (ambig)',
   'Male First Name (ambig)',
   'Last Name (ambig)',
   'First Name8 (NamePattern2)']},
```

pyDeid

Surrogate replacement

```
1 Frodo Baggins was born in The Shire on September 22, 2968 and
2 Peregrin Took was born in The Shire on April 8 2290.
```

- Example without surrogate replacement:

```
1 *** *** was born in *** on *** and
2 Peregrin Took was born in *** on ***.
```

- Example with surrogate replacement:

```
1 Saradoc Brandybuck was born in Rivendell on 1988/11/28 and
2 Peregrin Took was born in Rivendell on 1990/10/17.
```

pyDeid

- Re-identification

```
>> from pyDeid import deid_string, reid_string
>> original_string = 'Arwen Undómiel was born in Rivendell on March 1st, 24
>> phi, new_string = deid_string(original_string)

>> print(new_string)
'Jane Doe was born in London on 1976/6/3'

>> reid_string(new_string, phi)
'Arwen Undómiel was born in Rivendell on March 1st, 241'
```

pyDeid

- Visualization

```
>> from pyDeid import deid_string, display_deid
>> original_string = "Bilbo Baggins is a hobbit."
>> phi, _ = deid_string(example1)
>> display_deid(original_string, phi)
```

Bilbo **NAME** Baggins **NAME** is a hobbit.

pyDeid

- The `deid_string` and `reid_string` functions are designed for demonstration and testing.
- In settings where it is required to de-identify free text in bulk, we provide the `pyDeid` function.
- We can use this function on a test csv file.

pyDeid

- Working with files

```
>> pd.read_csv("my_awesome_data_file.csv")
genc_id note_id   note_text
1          Record 1  Aragorn II Ellesar is king of Gondor ...
2          Record 2  The Shire is located at 30 Hobbit St, Middle Earth ...
3          Record 3  Saruman the White and Gandalf the Grey are wizards ...
```

pyDeid

- Working with files

```
>> pd.read_csv("my_awesome_data_file.csv")

genc_id note_id   note_text
1          Record 1  Aragorn II Ellesar is king of Gondor
2          Record 2  The Shire is located at 30 Hobbit St, Middle Earth ...
3          Record 3  Saruman the White and Gandalf the Grey are wizards ...

>>> pyDeid(
...      original_file = 'my_awesome_data_file.csv',
...      note_varname = 'note_text',
...      encounter_id_varname = 'genc_id',
...      phi_output_file_type = 'json'
... )

Processing encounter 5: : 5it [00:00,  6.51it/s]

Diagnostics:
```

pyDeid

- Additional options

```
pyDeid(
    original_file: Union[str, pathlib.Path],
    note_varname: str,
    encounter_id_varname: str,
    new_file: Union[str, pathlib.Path, NoneType] = None,
    phi_output_file: Union[str, pathlib.Path, NoneType] = None,
    note_id_varname: Union[str, NoneType] = None,
    phi_output_file_type: Literal['json', 'csv'] = 'csv',
    custom_dr_first_names: Union[Set[str], NoneType] = None,
    custom_dr_last_names: Union[Set[str], NoneType] = None,
    custom_patient_first_names: Union[Set[str], NoneType] = None,
    custom_patient_last_names: Union[Set[str], NoneType] = None,
    verbose: bool = True,
    named_entity_recognition: bool = False,
    read_error_handling: str = None,
    **custom_regexes: str,
)
```

pyDeid

- Custom regular expressions
- There are cases where a particular data source might have some unique pattern that should be replaced.

```
>> example1 = "Your unique identifier for today's event is TMLS123456."
>> example1_phi, _ = deid_string(example1, ID='TMLS\d{6}')
>> display_deid(example1, example1_phi)
```

Your unique identifier for today's event is TMLS123456 **PHI** .

pyDeid

- Ensure the same de-identification is applied to the same person

```
>> example7 = """The Lord of the Rings film trilogy was directed by Pete Ja
...The screenplay was also written by Pete."""

>> example7_phi, example7_deid = deid_string(example7)
>> print(example7_deid)
The Lord of the Rings film trilogy was directed by Palma Kit.
...The screenplay was also written by Palma.
```

pyDeid

- Ensure that domain-specific terminology does not generate false positives.

```
>> example8 = "recommend Jackson-Pratt drain"
>> example8_phi, example8_deid = deid_string(example8)

>> print(example8_deid)
recommend Jackson-Pratt drain
```

pyDeid

- Maintain difference between dates

```
>> example9 = 'The Lord of the Rings trilogy was filmed in New Zealand for
... days from October 11, 1999 through December 22, 2000.'
>> example9_phi, example9_deid = deid_string(example9)

>> print(example9_deid)
>> The Lord of the Rings trilogy was filmed in New Zealand for 438 days fro
... August 30, 2007 through November 10, 2008.
```

Doctor names and patient names

- For some applications, the user may have access to a record of doctor and patient names associated with the clinical notes.
- When such lists are available, sensitivity ($\frac{TP}{TP+FN}$) can be improved by passing a `Set` containing these names to pyDeid.

Doctor names and patient names

```
>> example2 = "Bilbo Baggins is a hobbit and a patient at Shire General Hos
>> example2_phi, _ = deid_string(example1)
>> display_deid(example2, example2_phi)
```

Bilbo Baggins is a hobbit.

Doctor names and patient names

- By providing a custom list of patient or doctor names, we can do better:

```
>> example2_phi, _ = deid_string(
...     example2,
...     custom_patient_first_names={'Bilbo'},
...     custom_patient_last_names={'Baggins'}
... )

>> display_deid(example2, example2_phi)
```

Bilbo NAME Baggins NAME is a hobbit.

Doctor names and patient names

- Custom whitelists and blacklists

```
pydeid
| src\pyDeid
|- phi_types/
|- process_note/
|- wordlists/
|--- doctor_first_names.txt
|--- doctor_last_names.txt
|--- medical_phrases.txt
```

Regular Expression Rules for Names

- Simplified rule to detect names in text:

```
def titles(x, phi):
    specific_titles = ["MR", "MISTER", "MS"] # truncated list

    for title in specific_titles:
        for m in re.finditer(r'\b(' + title + r'\.( *))([A-Za-z\'\-]+)\b',
            potential_name = m.group(3)

            add_type(potential_name, "Name", phi)
```

Named Entity Recognition

- If we don't have access to a list of patient or doctor names, we can use **named entity recognition**.
- Named entity recognition is an "information retrieval" task to identify "structured information" in unstructured text. (Nadeau & Sekine, 2007)

```
>> example2_phi, _ = deid_string(example2, named_entity_recognition=True)

>> display_deid(example2, example2_phi)
```

Bilbo NAME Baggins NAME is a hobbit.

Named entity recognition

- Consider the case below where there is a name that is also an object.

```
>> example3 = "patient has to rely on walker to travel"
```

- This will to be recognized as PHI:

```
>> example3 = "patient has to rely on walker to travel"
>> example3_phi, _ = deid_string(example3)
>> display_deid(example3, example3_phi)
```

patient has to rely on walker to travel

Named entity recognition

- The example is somewhat ambiguous (it could be a grammar shortcut)
- If we capitalize `Walker`, it becomes more clear (to a human) that this is potentially referring to a person:

```
>> example4 = "patient has to rely on Walker to travel"
>> example4_phi, _ = deid_string(example4)
>> display_deid(example4, example4_phi)
```

patient has to rely on Walker to travel

- The regex-only approach fails to capture `Walker` as PHI.

Named entity recognition

- The regex-only approach failed to capture `Walker`.
- What if we used named entity recognition?

```
>> example4 = "patient has to rely on Walker to travel"
>> example4_phi, _ = deid_string(example4, named_entity_recognition=True)
>> display_deid(example4, example4_phi)
```

patient has to rely on Walker NAME to travel

- The named entity recognition approach is successful.

Named entity recognition

- The named entity recognition model uses `spaCy's` CNN-based NER model.
- There are alternative models such as the BERT-based NER model from `huggingface`:

```
>> from transformers import AutoTokenizer, AutoModelForTokenClassification
>> from transformers import pipeline

>> tokenizer = AutoTokenizer.from_pretrained("dslim/bert-base-NER")
>> model = AutoModelForTokenClassification.from_pretrained("dslim/bert-base

>> bert_ner = pipeline("ner", model=model, tokenizer=tokenizer)
```

Named entity recognition

```
1 import spacy
2
3 def display_ner(text, ner_result, title=None):
4     """Visualize NER with the help of SpaCy"""
5     ...
```

Named entity recognition

- Named entity recognition with the BERT NER.

```
>> bert_ner = pipeline("ner", model=model, tokenizer=tokenizer)

>> example3 = "patient has to rely on walker to travel"
>> display_ner(example3, bert_ner(example3))
```

patient has to rely on walker to travel

```
>> example4 = "patient has to rely on Walker to travel"
>> display_ner(example4, bert_ner(example4))
```

patient has to rely on Walker **B-PER** to travel

Named entity recognition

- The BERT-base (cased) model is able to catch "misspelled" names:

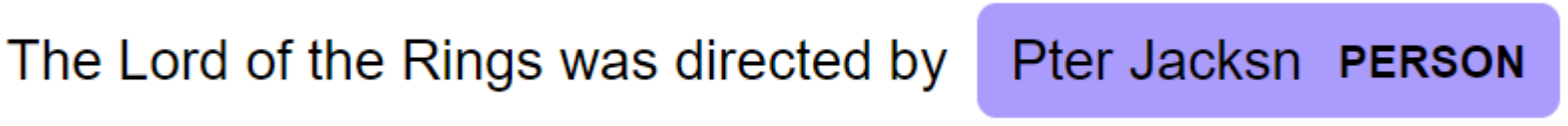
```
>> example5 = "The Lord of the Rings was directed by Pter Jacksn"
>> display_ner(example5, bert_ner(example5))
```

The Lord of the Rings **B-MISC** was directed by Pter Jacksn **B-PER**

Named entity recognition

- But `spaCy's` NER model also handles typos well:

```
>> example5 = "The Lord of the Rings was directed by Pter Jacksn"
>> from spacy import displacy
>> spacy_ner = spacy.load("en_core_web_sm")
>> displacy.render(spacy_ner(example5), style="ent")
```



Named entity recognition

- A regex-only approach fails to catch typos:

```
>> example5 = "The Lord of the Rings was directed by Pter Jacksn"
>> example5_phi, _ = deid_string(example5, named_entity_recognition=False)
>> display_deid(example5, example5_phi)
```

The Lord of the Rings was directed by Pter Jacksn

Named entity recognition

- There are still some issues with the CNN-based NER model
  - e.g., the distinction between Organizations and Persons

```
>> example6 = "Peregrin Took and Meriadoc Brandybuck are also hobbits"
>> displacy.render(spacy_ner(example6), style="ent")
```

Peregrin Took ORG and Meriadoc Brandybuck ORG are also hobbits

Named entity recognition

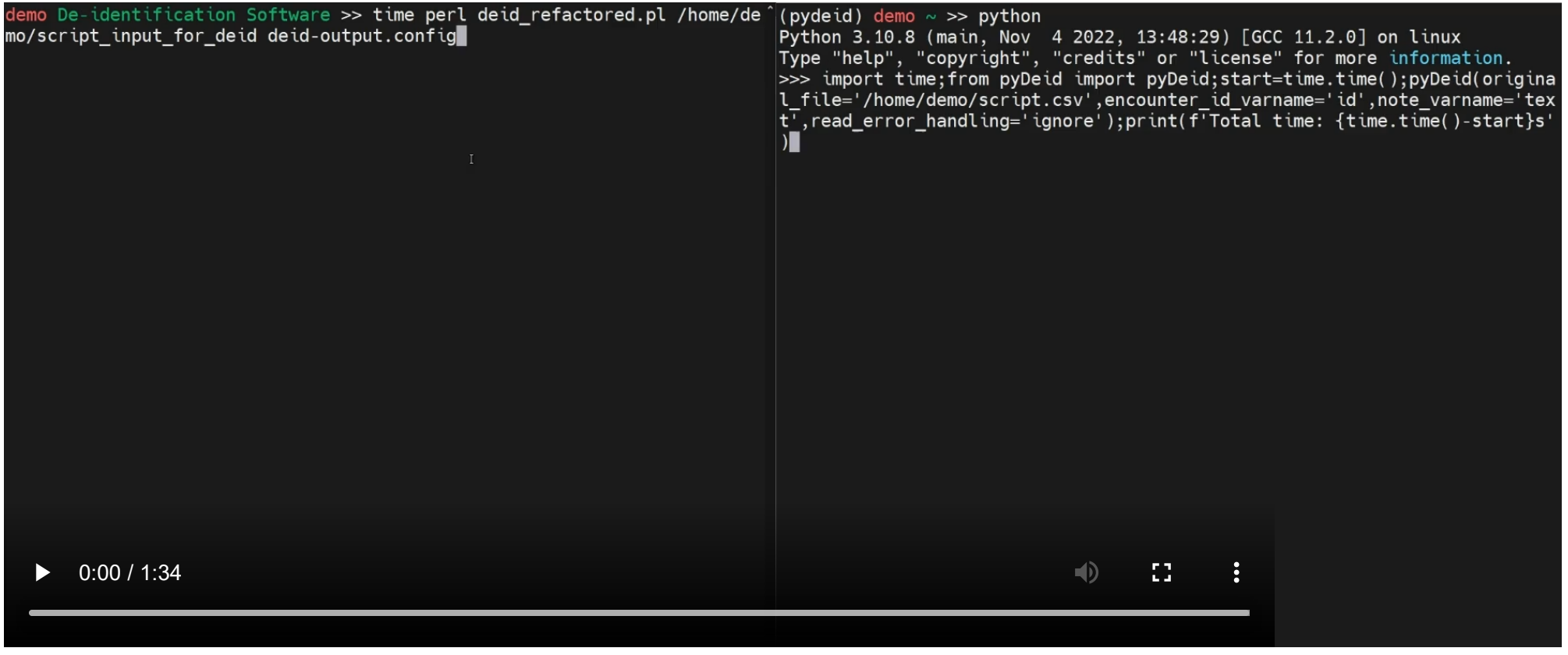
- The BERT-based NER model also struggles to distinguish between Organizations and Persons:

```
>> example6 = "Peregrin Took and Meriadoc Brandybuck are also hobbits"
>> display_ner(example6, bert_ner(example6))
```

Peregrin Took B-ORG and Meriadoc Brandybuck B-ORG are also hobbits

pydeid vs PhysioNet deid

- Lets compare time to de-identify all 3 scripts from the Lord of the Rings triology (53,971 words total)



pydeid vs PhysioNet deid

	chars/s	s/note	total time (s)
Physionet deid	3,208.28	31.44	94.31
pyDeid	29,093.58	3.43	10.40

Scalability

- High volume of notes:
  - 1.2M words per site per month
- Future work:
  - Distributed computing with Spark
  - Publish and open source

Conclusion

- Sharing data in healthcare is important
- The right tool depends on the context, problem, scope, etc
  - Lots of open source tools using different approaches
- De-identification of text data is the tip of the privacy iceberg

Resources, links, papers, etc.

- Unity Health Toronto: https://unityhealth.to/
- DSAA: https://unitynet.unity.local/departments-programs-services/corporate-services/data-science-and-advanced-analytics/
  - Blog: https://lks-chart.github.io/blog/
- GEMINI: https://www.geminimedicine.ca/
  - Download `pydeid` (coming soon!): github.com/GEMINI-Medicine

- What can we do with clinical notes?
  - "Assessment of Natural Language Processing Methods for Ascertaining the Expanded Disability Status Scale Score From the Electronic Health Records of Patients With Multiple Sclerosis: Algorithm Development and Validation Study." (Yang et al., 2022)

- Sharing data in healthcare is important
  - “Global healthcare fairness: We should be sharing more, not less, data.” (Seastedt et al., 2022)
  - “Dispelling the Myths Surrounding De-identification: Anonymization Remains a Strong Tool for Protecting Privacy.” (Information & Privacy Commissioner of Ontario, 2016)
  - “Deidentification Guidelines for Structured Data.” (Information & Privacy Commissioner of Ontario, 2016)
  - “Introduction to Data Sharing Rules”, (Information & Privacy Commissioner of Ontario, 2019)
  - “Concepts and Methods for De-identifying Clinical Trial Data”, (El Emam & Malin, 2014)

- Named Entity Recognition
  - Nadeau, D., & Sekine, S. (2007). “A survey of named entity recognition and classification.” (Nadeau & Sekine, 2007)
  - spaCy NER model
  - BERT NER model:
    - huggingface model: dslim/bert-base-NER
    - “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.” (Devlin et al., 2018)
    - “Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition.” (Sang & de Meulder, 2003)

- What do you need in a de-identification tool?
  - "Protected Health Information filter (Philter): accurately and securely de-identifying free-text clinical notes" (Norgeot et al., 2020)
  - "Comparing Rule-Based, Feature-Based and Deep Neural Methods for De-Identification of Dutch Medical Records." (Trienes et al., 2020)
  - "Using word embeddings to improve the privacy of clinical notes" (Abdalla et al., 2020)