Healthcare Applications with Natural Language Processing (NLP)

Raymond Ng

Director, Data Science Institute

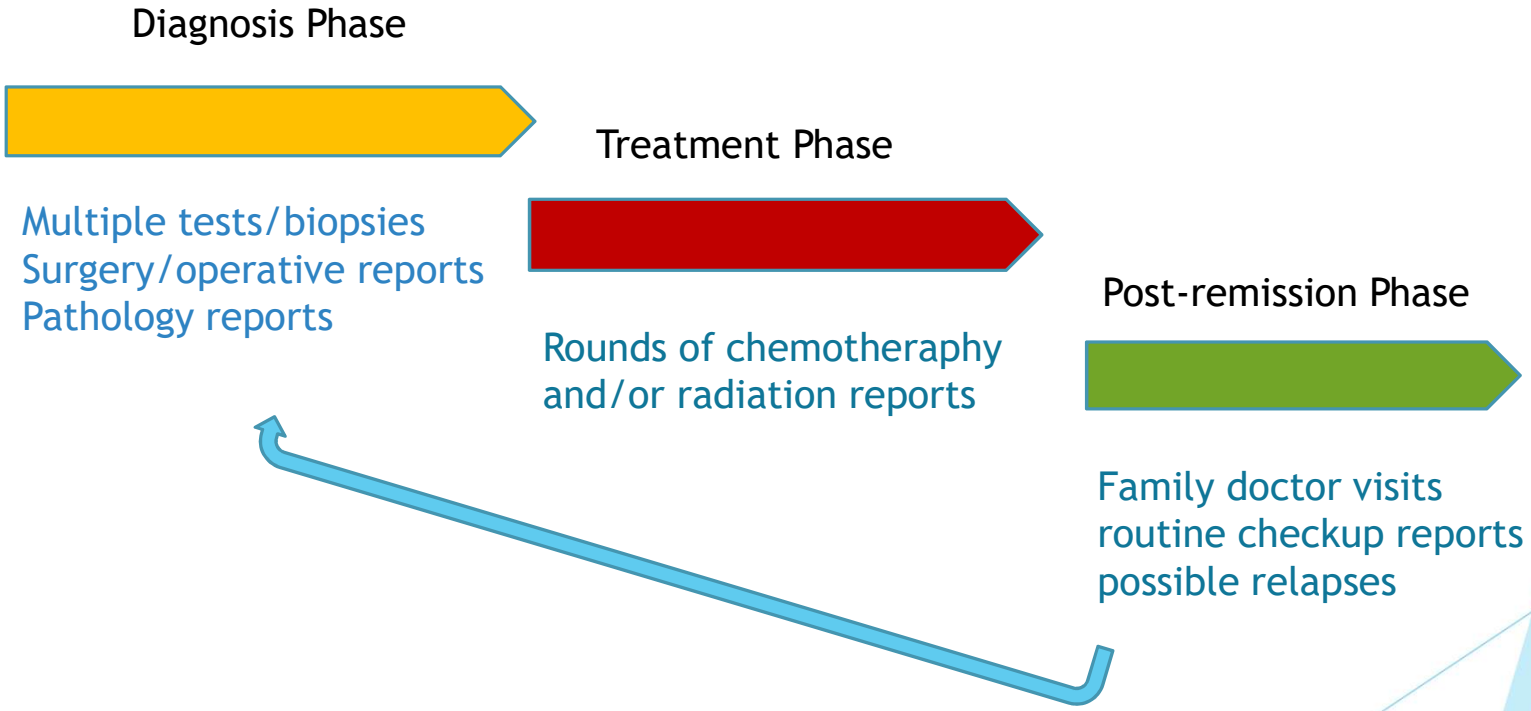
Canada Research Chair in Data Science and Analytics

Professor, Computer Science
University of British Columbia

Why NLP? First Answer: Gold mines

- Huge amounts of unstructured data
  1. “Official” documents, e.g., medical reports, clinical documents
  2. User generated content, e.g., blogs, text messages
- Machine Learning in health care still predominantly based on structured data, e.g. Electronic Health Records (EHR), insurance data
- Take home message: ***huge opportunities to use such unstructured data for better patient care***

Use Case for Clinical Documents: Cancer Care



Use Case for Official Documents: Cancer Care

- Current practice:
  - very siloed, different types of documents rarely linked together
    - Used mainly to make the next "local" decisions
    - Key findings over the whole journey not consolidated "holistically"
  - Documents read/interpreted only by humans
    - Lack of health care resources means significant time delay, e.g., many months
- Our ongoing work is to use NLP and machine learning to automatically extract "*fields of interest*" from clinical documents
  - Improving detection of reportable cancer
  - Identifying diagnostic groups
  - Predicting treatment response
  - Stratifying risks for relapses

E.g., A Breast Cancer Surgical Report:

staging, location, size, negation, etc.

Final Diagnosis ::
A. Right mastectomy including nipple and skin showing:
1. Two loci of recurrent invasive ductal carcinoma of primary breast derivation.
2. Both are located in the upper one-half of the specimen in the vicinity of prior surgery with one deposit designated at the 11:30 o'clock position

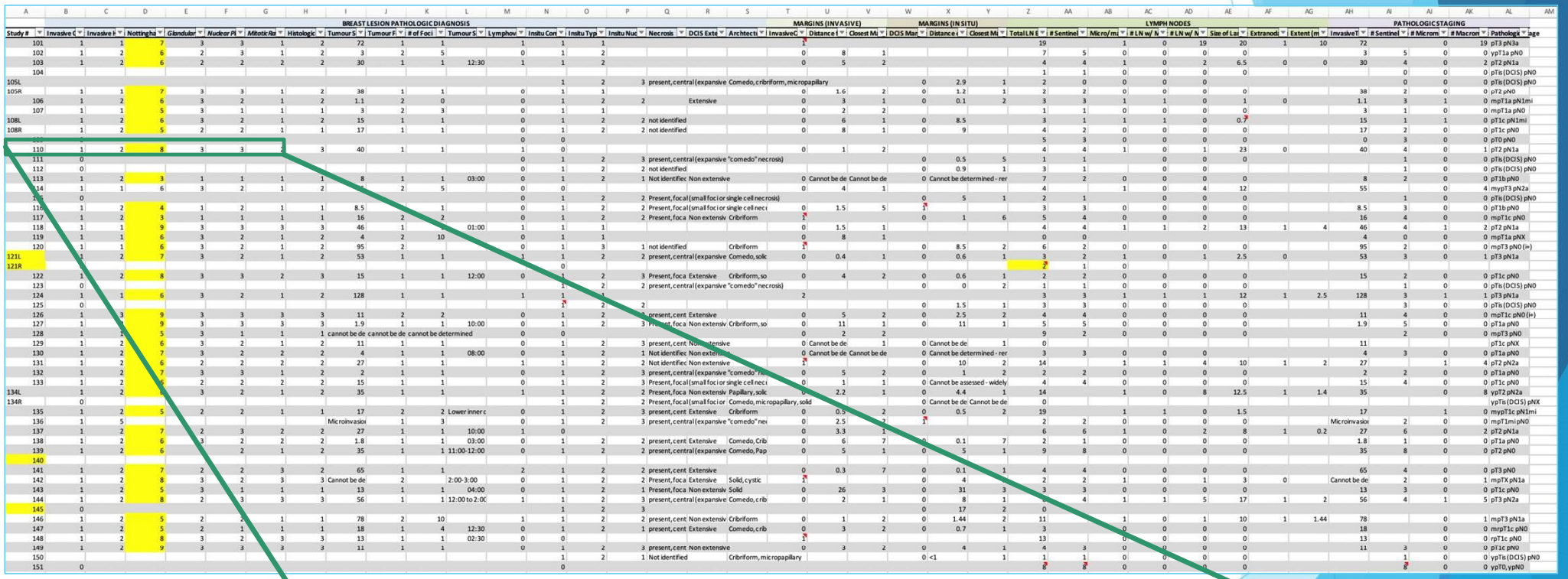
measuring 1.5 x 1.2 cm with the other at roughly the 10 o'clock position measuring 1.2 x 1 cm.
3. These deposits are separate from one another by 3 cm.
4. Both are very similar histologically and each have pleomorphic somewhat apocrine nuclei (score 3); are 50% gland forming (score 2) and show a variable mitotic rate focally above 7/mm2 (score 3) and thus overall Nottingham score is 8/9, grade 3.
5. Prognostic marker status: previously performed on both nodules through Lion's Gate Hospital on prior needle core biopsies (LS19-1426) showing:
   a. Estrogen receptor status positive, Allred score 8/8.
   b. Progesterone receptor negative, Allred score 0/8.
   c. HER2neu by immunohistochemistry negative.
   d. These results are not similar to the original tumor of 2016 (LS16-13442).
   -No lymphovascular invasion identified (extensive retraction artefact around tumoral nests makes assessment difficult)
6. Both deposits of invasive carcinoma harbor a small amount of intratumoral intraductal carcinoma with high-grade apocrine nuclei with no comedo necrosis.

//emr p ex a ca/emr saac/home/PathNet/pathnetHTML aspx?c...193&message d=2&msh d=30803&obx d=91652&pt d=1460&showsess on=1 Page

ts For VCH 2020 06 11 12

7. Margins:
   a. 11:30 tumor lies at closest approach 2 mm from the blue anterior margin (block A3) - other margins widely negative.
   b. Separate deposit at roughly 10 o'clock involves the anterior margin via a 'tongue' 2 mm in cross dimension (tumour found within cautery at blue anterior ink) and at closest approach is 2 mm from the deep margin (block A7/A8) with other margins widely negative
8. No nipple involvement and all other four quadrants extensively sectioned show no in-situ carcinoma or infiltrating carcinoma.

Output: an Extracted Table of Fol's



Study #	Invasive Carcinoma	Invasive Histologic Type	Nottingham Score	*Glandular Differentiation*	*Nuclear Pleomorphism*	*Mitotic Rate*
110	1	2	8	3	3	2

An Incomplete List of Fields of Interest

Mastectomy

- Study #
- Invasive Carcinoma
- Invasive Histologic Type
- Nottingham Score
- Glandular Differentiation
- Nuclear Pleomorphism
- Mitotic Rate
- Histologic Grade
- Tumour Size (mm)
- Tumour Focality
- # of Foci
- Lymphovascular Invasion
- Tumour Site
- Insitu Component
- Insitu Type
- Insitu Nuclear Grade
- Necrosis
- DCIS Extent
- Archtectural Patterns

Margins

- Invasive Carcinoma Margins
- Distance from Closest Margin
- Closest Margin

- DCIS Margins
- Distance of DCIS from
- Closest Margin (mm)
- Closest Margin DCIS

Lymph Nodes

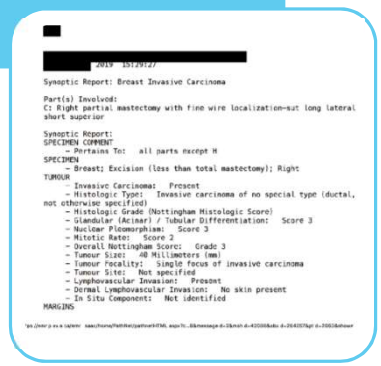
- Total LN Examined
- # Sentinel LN Examined
- Micro/macro metastasis
- # LN w/ Micrometastasis
- # LN w/ Macrometastasis
- Size of Largest Macrometastasis Deposit
- Extranodal Extension
- Extent (mm)

Pathologic Staging

- Invasive Tumour Size (mm)
- # Sentinel Nodes Examined
- # Micrometastatic Nodes
- # Macrometastatic Nodes
- Pathologic Stage

"Automated medical chart review for breast cancer outcomes research: a novel natural language processing extraction system" (BMC May 2022)

Input PDF

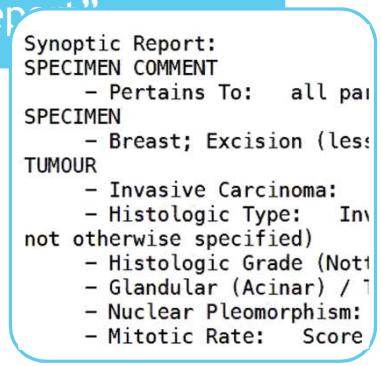


Synoptic Report: Breast Invasive Carcinoma

Part(s) Involved:
C: Right partial mastectomy with fine wire localization-sut long lateral short superior

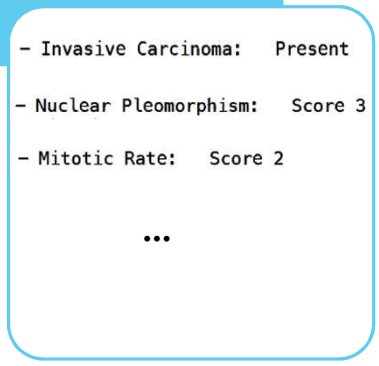
Synoptic Report:
SPECIMEN COMMENT
- Pertains To: all parts except H
SPECIMEN
- Breast; Excision (less than total mastectomy); Right
TUMOUR
- Invasive Carcinoma: Present
- Histologic Type: Invasive carcinoma of no special type (ductal, not otherwise specified)
- Histologic Grade (Nottingham Histologic Score)
- Glandular (Acinar) / Tubular Differentiation: Score 3
- Nuclear Pleomorphism: Score 3
- Mitotic Rate: Score 2
- Overall Nottingham Score: Grade 3
- Tumour Size: 40 Millimeters (mm)
- Tumour Focality: Single focus of invasive carcinoma
- Tumour Site: Not specified
- Lymphovascular Invasion: Present
- Dermal Lymphovascular Invasion: No skin present
- In Situ Component: Not identified
MARGINS

Search for "Synoptic Report"



Synoptic Report:
SPECIMEN COMMENT
- Pertains To: all pa
SPECIMEN
- Breast; Excision (les
TUMOUR
- Invasive Carcinoma:
- Histologic Type: In
not otherwise specified)
- Histologic Grade (Not
- Glandular (Acinar) /
- Nuclear Pleomorphism:
- Mitotic Rate: Score

Isolate each bullet-point



- Invasive Carcinoma: Present
- Nuclear Pleomorphism: Score 3
- Mitotic Rate: Score 2
...

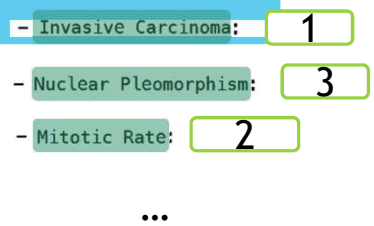
Extract data

- Nuclear Pleomorphism: Score 3
- Invasive Carcinoma: Present
- Mitotic Rate: Score 2
...

Save raw text to Excel

	A	B	C	D
	Study #	Invasive Carcinoma	Nuclear Pleo-morphism	Mitotic Rate
	110	Present	Score 3	Score 2

Encode data



- Invasive Carcinoma: 1
- Nuclear Pleomorphism: 3
- Mitotic Rate: 2
...

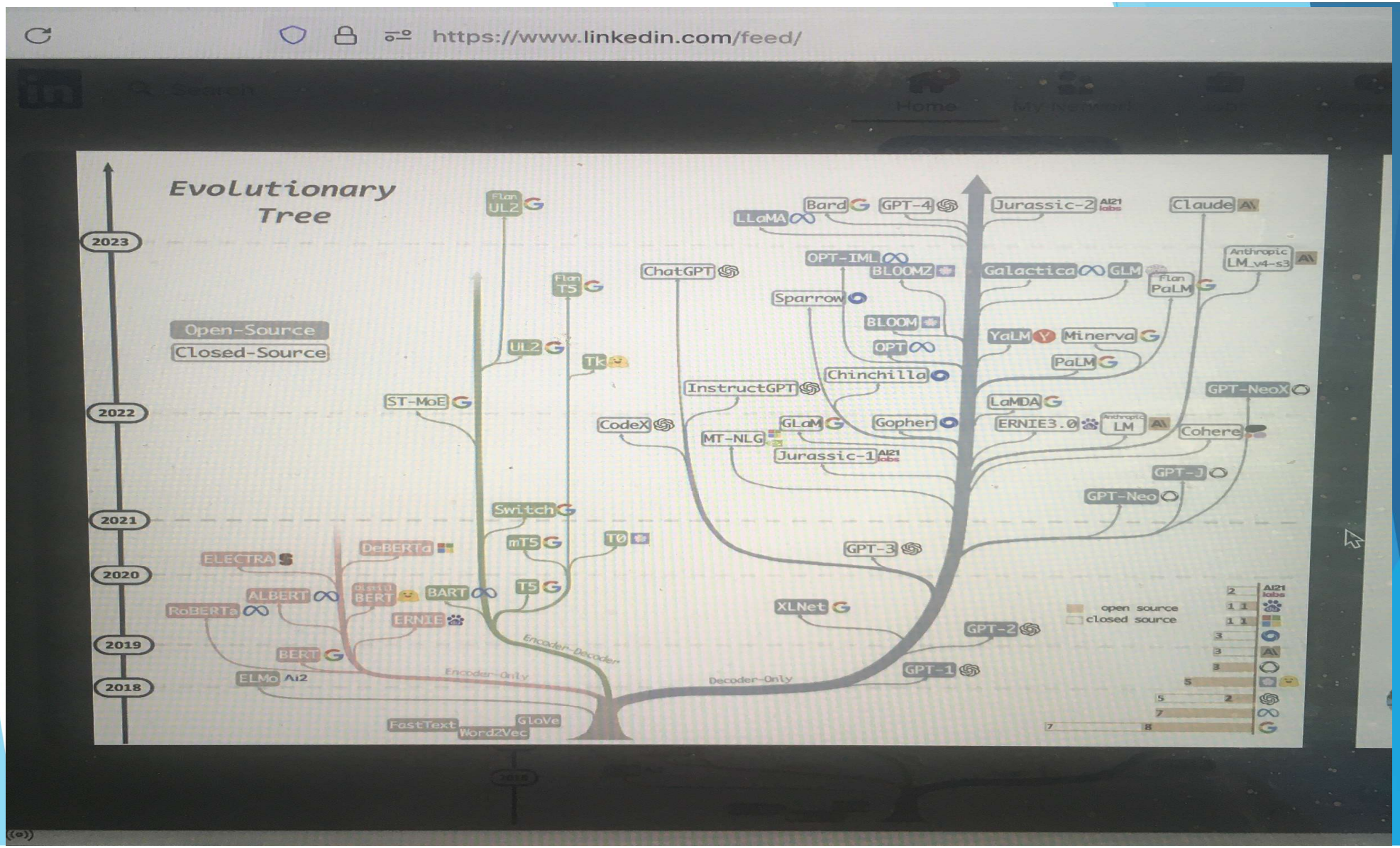
Save encodings to Excel

	A	B	C	D
	Study #	Invasive Carcinoma	Nuclear Pleo-morphism	Mitotic Rate
	110	1	3	2

...

Why NLP? Second Answer: Great Advances

- Pre-trained language models - biggest advances in NLP this decade
  - Trained with a large dataset while remaining agnostic to the specific tasks they will be employed on
  - E.g., BERT: created by Google with from English Wikipedia with 2,500M words
    - NLP -> **Natural Language Understanding**
  - Many variants, e.g., BioBERT, PubMed BERT, RoBERTa
  - Later models, e.g., T5, GPT2, GPT3, and ChatGPT
- Designed to be “fine-tunable” with specific tasks and domains, e.g., questions and answers


https://www.linkedin.com/feed/
Evolutionary Tree
2023
2022
2021
2020
2019
2018
Open-Source
Closed-Source
Flan UL2
Bard
GPT-4
Jurassic-2
Claude
LLaMA
OPT-IML
BLOOMZ
ChatGPT
Galactica
GLM
Anthropic LM_v4-s3
Flan T5
Sparrow
Flan PaLM
BLOOM
YaLM
Minerva
UL2
OPT
Tk
PaLM
Chinchilla
InstructGPT
GPT-NeoX
ST-MoE
LaMDA
CodeX
GLaM
Gopher
ERNIE3.0
Anthropic LM
Cohere
MT-NLG
Jurassic-1
GPT-J
GPT-Neo
Switch
mT5
T0
GPT-3
ELECTRA
DeBERTa
ALBERT
distil BERT
BART
T5
RoBERTa
ERNIE
XLNet
GPT-2
BERT
Encoder-Decoder
Encoder-Only
Decoder-Only
GPT-1
ELMo
AI2
FastText
GloVe
Word2Vec
open source
closed source

Why NLP? BERT Q/A examples

- From: //huggingface.co/tasks/question-answering
- E.g., text: "*I am Sarah and Vancouver is my home*"
  - Q1: "*what is my name*"
  - Ans1: "Sarah"
  - Q2: "*where do I live*"
  - Ans2: "Vancouver"
- E.g., text: "*The Amazon rainforest, also known in English as Amazonia or the Amazon Jungle*"
  - Q1: "*Which name is also used to describe the Amazon rainforest*"
  - Ans: "Amazonia"
- We use it for healthcare applications and mineral mining governance (environmental waste management)

What about User Generated Content (UGC)?

- Clinical documents written by clinicians and healthcare professionals
- What about listening to the patients, their families and care-givers?
  - Their opinions, experiences, needs, feelings, mood, etc.
- How is UGC different from formal documents?
  - Diverse backgrounds
  - Diverse writing styles: use of words, length, may not even be grammatically correct
  - More subjective
  - Different genre: forums, chats, conversations, blogs

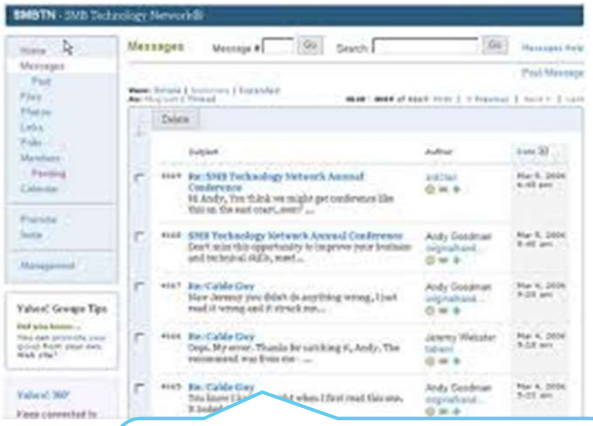
Listening to the Patients

Topic modeling
What are patients talking about? E.g., diet, treatment, doctors, …

Needs detection
What needs are patients seeking?
e.g., health information, social support, …

Dialog acts
Why are patients talking about certain topics?
e.g., explain information on a diet, ask questions on skin rash, sharing emotions on a treatment

Patients' conversations



Discourse coherence
Is what the patient says coherent?

Text Analysis for Chronic Disease Management

- Topic modeling
  - Research task: topic modeling with an ontology
  - Data: mDAWN social media discussions
- Needs detection
  - Research task: needs prediction
  - Data: American Cancer Society online discussion forum data
- Dialog acts
  - Research task: dialog act prediction
  - Data: American Cancer Society online forum data
- Discourse coherence
  - Research task: dementia detection
  - Dementia datasets (speech and text)

Online Discussion Forums, e.g., American Cancer Society

- Understand their needs as a first step to recommending interventions
- But Manual identification of needs too labor intensive and time consuming
- Develop automated algorithms to classify types of needs
  - Data consists of: 52,000+ posts (2006-2016)
  - Collected from the Cancer Survivors Network online peer-support forum (http://csn.cancer.org)

"Neural Prediction of Patient Needs in an Ovarian Cancer Online Discussion Forum" (Canadian AI, 2019)

- **Physical needs:** *"I have finished the 6 rounds of carbo/taxol and am slowly recovering from the assault of the treatment. Has anyone experienced joint pain after tx? I've noticed increasing pain in both hips and recently knee pain."*

- Cross validation accuracy: 85%

- **Emotional needs:** *"I will have 6-8 months of chemo. I am so scared. I do not know what to expect. I am having a hard time dealing with this and am going to attend a support meeting in April. I having a difficult time staying positive and upbeat. I have read such amazing stories on here I hoping that this board will help me as well..((((HUGS TO ALL))))"*

- Cross-validation accuracy: 80%

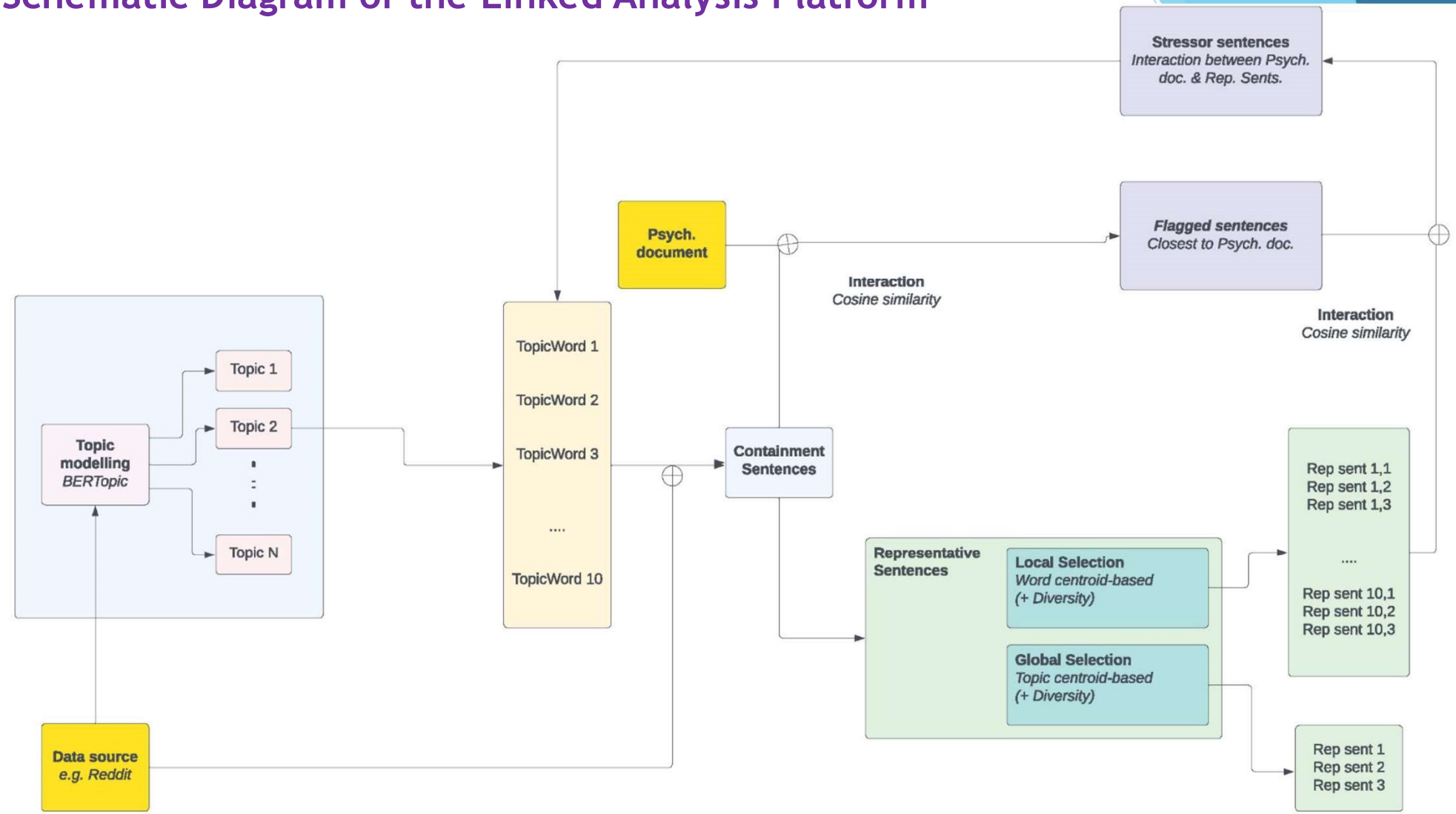
Another Use Case: Mental Health and Psychiatry

- Part 1, Clinical documents: in psychiatry, even clinical documents, e.g., psychiatric assessments, are highly unstructured
- Part 2, User generated content: social media posts
  - To monitor at-risk adolescents through their instant messaging content (because over 80% of all youths are heavy users of social media)
  - Develop a recommendation system for early intervention
- Dataset: 1,000 youths admitted to hospitals through emergency room due to self-harm: their admission notes, discharge summaries, and their social media posts 6 months prior to admissions
- One predictive model we are building: identify high-risk re-admission cases

A Platform for Online Psychiatric Analysis on Social Media Posts

- A Health Canada funded project to monitor the mental health of university students across Canada, eg., UBC, U of Toronto, etc.
- Enrolled 35,000 university students across Canada
- UBC students alone generated 300,000+ Reddit posts between 2020 and 2022
- Selected questions of interest:
  - What do they talk about?
  - What stress them out?
  - How does specific stress change over time?
  - How does specific stress differ across campuses/locations?

Schematic Diagram of the Linked Analysis Platform



Two Final Remarks

- Multi-lingual issue
  - Most NLP research driven by English corpora
  - One way to try a different language X is by automatic translation of documents in X to English, and apply the English models
  - A longer-term way is to apply transfer learning to English large language models to build models from documents in language X
- Even though we talk about written text so far, what about speech
  - Huge amounts of data collected by speech technologies, e.g., Siri for Apple, Alexa for Amazon
  - One way is to automatically transcribe speech to text and apply NLP-based models

Thank You!

rng@cs.ubc.ca
//dsi.ubc.ca

Future Work in Mental Health

- Adapt the model to monitor at-risk adolescents through their instant messaging content (because over 80% of all youths are heavy users of social media)
  - Develop a recommendation system for early intervention
- Apply transfer learning to build models for other populations:
  - Cancer patients
  - Patients with serious chronic conditions who stay at home
  - Seniors
  - Isolated individuals, e.g., covid-19