UCSF
University of California
San Francisco

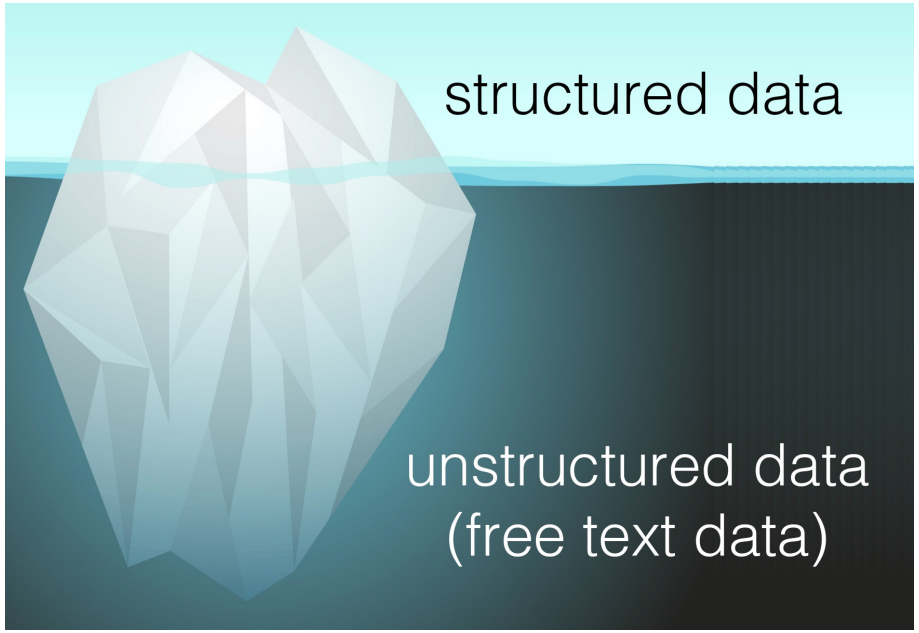
Obtaining Deeper Insights From EMR Clinical Notes for the Longitudinal Management of Brain Mets

Hui Lin, PhD
Department of Radiation Oncology
University of California San Francisco
July 12th, 2022

EMR: Significance and Challenges

- An EMR allows the electronic entry, storage, and maintenance of digital medical data. EHR contains the patient's records from doctors and includes demographics, test results, medical history, history of present illness (HPI), and medications. EMRs are part of EHRs and contain the following:
  - Patient registration, billing, preventive screenings, or checkups
  - Patient appointment and scheduling
  - Tracking patient data over time
  - Monitoring and improving overall quality of care

- Structured vs. Unstructured data



Courtesy of https://doi.org/10.1155/2020/5471849, https://project-emerse.org/data_in_free_text.html

What is Natural Language Processing?

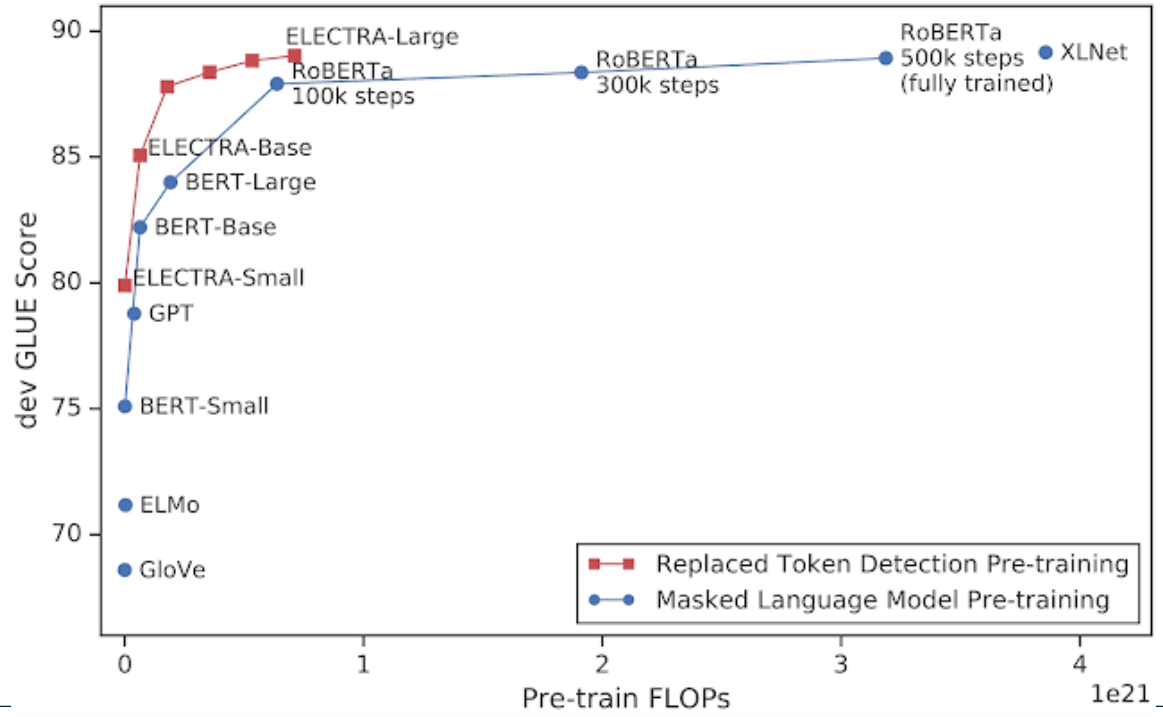
- Natural Language Processing (NLP): Exploiting patterns and extracting key information from **unstructured** clinical texts

- Major approaches:
  - Rule-based: Emulating rules informed by the structure of language
  - Statistical/Machine Learning: Generating probabilistic rules by using large text corpora

- EHR notes can be tough to work with since:
  - Corpus assembly challenge
  - Linguistic complexity
  - Heterogeneous structure

- Deep learning has revolutionized NLP with a new approach
  - **Pre-train** a general language model with massive, diverse corpora of unlabeled texts using self-supervised learning
  - **Transfer Learning** to adapt the language model to a target domain
  - Use the pre-trained weights and only a **small number of labeled data** to train the model for a downstream task

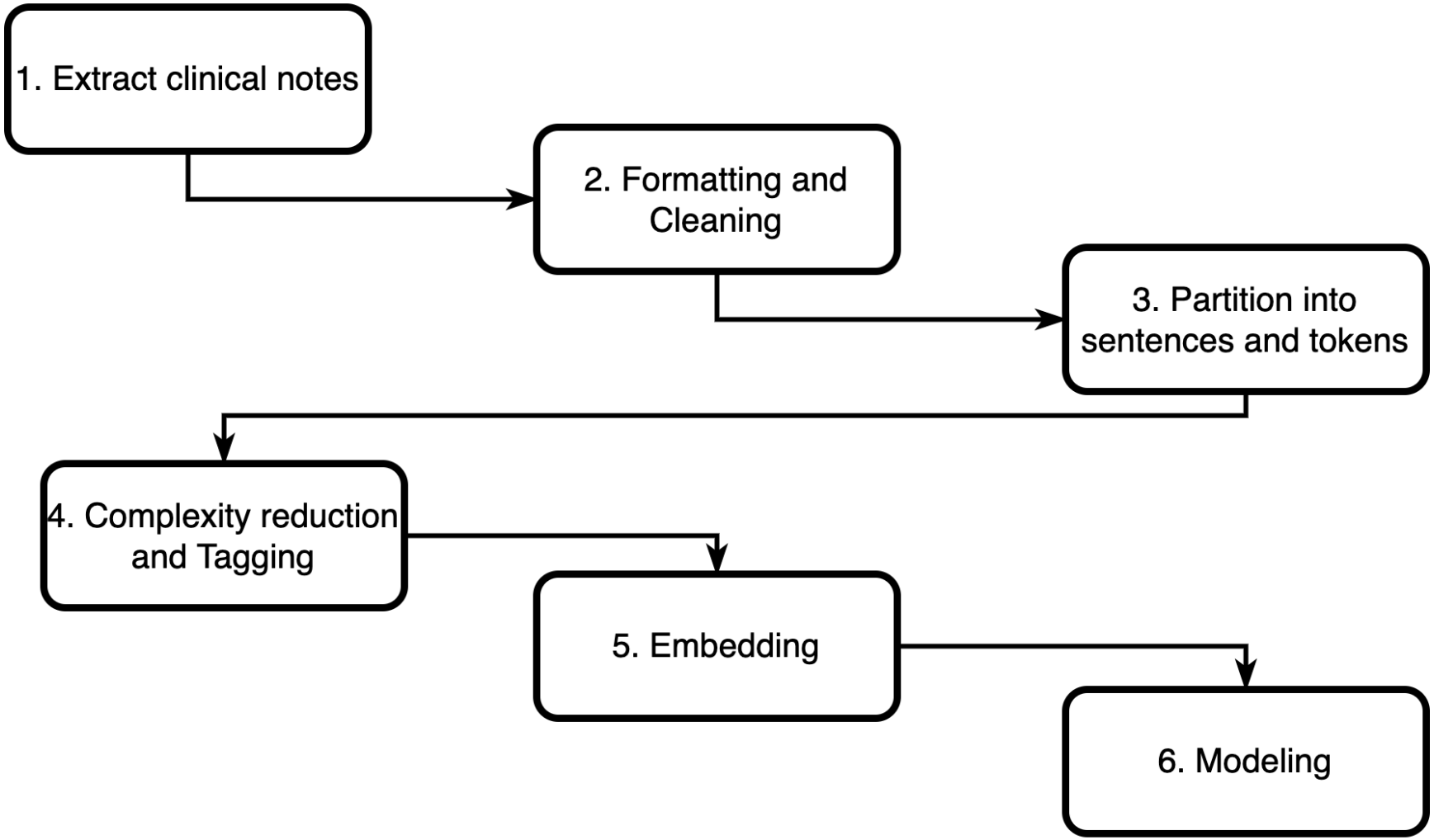
- Bidirectional Encoder Representations from Transformers (BERT) is one of the most commonly used deep learning models in NLP



Courtesy of google blog: https://www.googblogs.com/more-efficient-nlp-model-pre-training-with-electra/

UCSF

General NLP Pipelines



Step 1: Extraction

- To filter through unstructured clinical notes, extract relevant information and convert into a consistent format that is editable and accessible

- The first step in the pipeline is programmatically extracting the free text information using SQL, API calls or some other means of harvesting texts

- Best to store the extracted data in memory, or an easy to ingest text format (csv/json/pickle) along with other structured data

Step 1: Extraction

```
import PyODBC
connection = pyodbc.connect("DRIVER=;SERVER=")
cursor = connect.cursor()
sql_cmd = 'select id, note_text from admission where id = 22'
data_obj = cursor.execute(sql_cmd)

COPY admissions
FROM 'data/mimic-iii-clinical-database-1.4/ADMISSIONS.csv'
WITH (FORMAT csv,   HEADER true);

CREATE TABLE patients (
    row_id int,
    subject_id DECIMAL,
    gender VARCHAR,
    dob TIMESTAMP,
    dod TIMESTAMP,
    dod_hosp TIMESTAMP,
    dod_ssn TIMESTAMP,
    expire_flag BOOLEAN
);

COPY patients
FROM 'data/mimic-iii-clinical-database-1.4/PATIENTS.csv'
WITH (FORMAT csv,   HEADER true);

CREATE MATERIALIZED VIEW adm_details as
(
    select p.subject_id, p.gender, p.dob, p.dod, hadm_id, admittime, dischtime, admission_type,
    insurance, marital_status, ethnicity, hospital_expire_flag, has_chartevents_data
    from admissions adm
    join patients p
    on adm.subject_id = p.subject_id
)

select * from adm_details;

Copy (select * from adm_details) To 'adm_details.csv' With CSV DELIMITER ',' HEADER;
```

subject_id	gender	dob	dod	hadm_id	admittime	dischtime	admission_type	insurance	marital_status	ethnicity	hospital_expire_flag	has_chartevents_
22	F	2131-05-07 00:00:00	NaN	165315	2196-04-09 12:26:00	2196-04-10 15:54:00	EMERGENCY	Private	MARRIED	WHITE	0	
23	M	2082-07-17 00:00:00	NaN	152223	2153-09-03 07:15:00	2153-09-08 19:10:00	ELECTIVE	Medicare	MARRIED	WHITE	0	
23	M	2082-07-17 00:00:00	NaN	124321	2157-10-18 19:34:00	2157-10-25 14:00:00	EMERGENCY	Medicare	MARRIED	WHITE	0	
24	M	2100-05-31 00:00:00	NaN	161859	2139-06-06 16:14:00	2139-06-09 12:48:00	EMERGENCY	Private	SINGLE	WHITE	0	
25	M	2101-11-21 00:00:00	NaN	129635	2160-11-02 02:06:00	2160-11-05 14:55:00	EMERGENCY	Private	MARRIED	WHITE	0	
...	...	...	...	...	...	...	...	...	...	...	...	

Exemplar notes from: MIMIC-III (https://doi.org/10.13026/C2XW26)

UCSF

Step 2: Formatting and cleaning

- To remove redundant information or problems in the notes, define a search pattern and then determine how to correct
- Considerations for unstructured clinical notes:
  - Abbreviations
  - Implied tables
  - Stray punctuation and spaces
  - Omitted punctuation
- Potential corrections in this step:
  - Normalizing punctuation and spacing
  - Transformation to lowercase letters
  - Stemming
- Text cleaning is laborious without automated pattern matching:
  - One common approach is regular expressions (regex)
  - Every major programming language has support for them (Python: re library)

Step 2: Formatting and cleaning

```
'8 am ct head w o contrast clip reason. Please eval for melanoma mets prior to anticoagulating for ***** admitting di
agnosis acute coronary syndrome medical condition, 79 year old man with h o melanoma non hodgkin s lymphoma here w ch
est pain ruling in for mi reason for this examination. Eval for melanoma mets prior to anticoagulating for mi no cont
raindications for iv contrast pfi report no evidence of acute intracranial hemorrhage seen although there is no evide
nce of edema shift of normally midline structures or effacement of the basal cisterns to suggest a brain mass a contr
ast study or mri would be recommended if there remains concern for intracranial metastatic lesion sinus mucosal disea
se involving ethmoid and sphenoid air cells'
```

```python
import re
import inflect
# remove unicode marks
notes = re.sub(r'!\s*', '', notes)

# remove de-identified brackets
notes = re.sub('\*', '', notes)

# replace number with words
notes = re.sub(r"\b\d+\b", lambda m: inflect.engine().number_to_words(m.group()), notes)
notes
```

```
'eight am ct head w o contrast clip reason. Please eval for melanoma mets prior to anticoagulating for  admitting dia
gnosis acute coronary syndrome medical condition, seventy-nine year old man with h o melanoma non hodgkin s lymphoma
here w chest pain ruling in for mi reason for this examination. Eval for melanoma mets prior to anticoagulating for m
i no contraindications for iv contrast pfi report no evidence of acute intracranial hemorrhage seen although there is
no evidence of edema shift of normally midline structures or effacement of the basal cisterns to suggest a brain mass
a contrast study or mri would be recommended if there remains concern for intracranial metastatic lesion sinus mucosa
l disease involving ethmoid and sphenoid air cells'
```

Step 3: Sentence segmentation

- To identify sentence boundaries between words in different sentences

- Identifying where setences end can be challenging in clinical notes – written in short form without punctuations:
  - *“4. Decadron taper down to 2 mg p.o. b.i.d. over a week's time. Currently Decadron is at 4 mg p.o. q6hours.5 Heparin subcutaneous 5000 units b.i.d.”*

- Sentence segmentation is a combination of rule-based and statistical processes

Step 3: Sentence segmentation

```
import spacy
# load the English model
seg_model = spacy.load('en_core_web_sm')
# define the document model
doc = seg_model(notes)
# list the segmented sentences
for i,s in enumerate(doc.sents):
    print (str(i) + '-->' + s.text.strip())
```

```
0-->eight am ct head w o contrast clip reason.
1-->Please eval for melanoma mets prior to anticoagulating for  admitting diagnosis acute coronary syndrome medical c
ondition, seventy-nine year old man with h o melanoma non hodgkin s lymphoma here w chest pain ruling in for mi reaso
n for this examination.
2-->Eval for melanoma mets prior to anticoagulating for mi no contraindications for iv contrast pfi report no evidenc
e of acute intracranial hemorrhage seen although there is no evidence of edema shift of normally midline structures o
r effacement of the basal cisterns to suggest a brain mass a contrast study or mri would be recommended if there rema
ins concern for intracranial metastatic lesion sinus mucosal disease involving ethmoid and sphenoid air cells
```

Step 4: Tokenization

- To further break each sentence down as a collection of tokens

- A token can be:
  - a word or sub-word structure (uni-gram)
    - ”tumor” (uni-gram)
    - #itis (sub-word token)
  - a collection of n words (n-gram)
    - ”brain tumor” (bi-gram)
    - “right frontal anaplastic” (tri-gram)

Step 4: Tokenization

- Regular expression-based:

```
import nltk
tokens = nltk.tokenize.word_tokenize(notes)
```

```
['8', 'am', 'ct', 'head', 'w', 'o', 'contrast', 'clip', 'reason', '.', 'Please', 'eval', 'for', 'melanoma', 'mets',
'prior', 'to', 'anticoagulating', 'for', '*****', 'admitting', 'diagnosis', 'acute', 'coronary', 'syndrome', 'medica
l', 'condition', ',', '79', 'year', 'old', 'man', 'with', 'h', 'o', 'melanoma', 'non', 'hodgkin', 's', 'lymphoma', 'h
ere', 'w', 'chest', 'pain', 'ruling', 'in', 'for', 'mi', 'reason', 'for', 'this', 'examination', '.', 'Eval', 'for',
'melanoma', 'mets', 'prior', 'to', 'anticoagulating', 'for', 'mi', 'no', 'contraindications', 'for', 'iv', 'contras
t', 'pfi', 'report', 'no', 'evidence', 'of', 'acute', 'intracranial', 'hemorrhage', 'seen', 'although', 'there', 'i
s', 'no', 'evidence', 'of', 'edema', 'shift', 'of', 'normally', 'midline', 'structures', 'or', 'effacement', 'of', 't
he', 'basal', 'cisterns', 'to', 'suggest', 'a', 'brain', 'mass', 'a', 'contrast', 'study', 'or', 'mri', 'would', 'b
e', 'recommended', 'if', 'there', 'remains', 'concern', 'for', 'intracranial', 'metastatic', 'lesion', 'sinus', 'muco
sal', 'disease', 'involving', 'ethmoid', 'and', 'sphenoid', 'air', 'cells']
```

- Deep learning-based:

```
from pytorch_pretrained_bert.tokenization import BertTokenizer
tokenizer = BertTokenizer.from_pretrained('bert-base-uncased')
tokens = tokenizer.tokenize(notes)
```

```
['8', 'am', 'ct', 'head', 'w', 'o', 'contrast', 'clip', 'reason', '.', 'please', 'eva', '##l', 'for', 'mel', '##ano',
'##ma', 'mets', 'prior', 'to', 'anti', '##co', '##ag', '##ulating', 'for', '*', '*', '*', '*', '*', 'admitting', 'dia
gnosis', 'acute', 'corona', '##ry', 'syndrome', 'medical', 'condition', ',', '79', 'year', 'old', 'man', 'with', 'h',
'o', 'mel', '##ano', '##ma', 'non', 'ho', '##d', '##g', '##kin', 's', 'l', '##ym', '##ph', '##oma', 'here', 'w', 'che
st', 'pain', 'ruling', 'in', 'for', 'mi', 'reason', 'for', 'this', 'examination', '.', 'eva', '##l', 'for', 'mel', '#
#ano', '##ma', 'mets', 'prior', 'to', 'anti', '##co', '##ag', '##ulating', 'for', 'mi', 'no', 'contra', '##ind', '##i
cation', '##s', 'for', 'iv', 'contrast', 'p', '##fi', 'report', 'no', 'evidence', 'of', 'acute', 'intra', '##cr', '##
anial', 'hem', '##or', '##rh', '##age', 'seen', 'although', 'there', 'is', 'no', 'evidence', 'of', 'ed', '##ema', 'sh
ift', 'of', 'normally', 'mid', '##line', 'structures', 'or', 'e', '##ffa', '##ce', '##ment', 'of', 'the', 'basal', 'c
is', '##tern', '##s', 'to', 'suggest', 'a', 'brain', 'mass', 'a', 'contrast', 'study', 'or', 'mri', 'would', 'be', 'r
ecommended', 'if', 'there', 'remains', 'concern', 'for', 'intra', '##cr', '##anial', 'meta', '##static', 'les', '##io
n', 'sin', '##us', 'mu', '##cos', '##al', 'disease', 'involving', 'et', '##hm', '##oid', 'and', 'sp', '##hen', '##oi
d', 'air', 'cells']
```

Optional: Complexity reduction and Tagging

- To reduce the overall complexity, as the original set of tokens may have redundancy and various morphological forms of the same concept

- Complexity reduction techniques:
  - Spelling and abbreviation correction
  - Stemming: converting words to their root forms
  - Stop word removal: get rid of common words like ’a’, ‘the’, ’to’

- POS Tagging is to identify the grammatical categories of each token

- POS Tagging can be done in multiple ways:
  - Token by token using dictionary of the word’s POS
  - Statistical approaches using entire sentence

Summary of Pipeline in Practice

1. Notes extraction from Clarity
2. Formatting and cleaning each note using Python string functions and regular expressions

60%

3. Sentence segmentation, tokenize the notes into n-grams
4. Develop the vocabulary with all words used within the notes corpora
5. Build a model for your downstream task with the input of your NLP model (Tf-idf, Word2Vec, BERT etc.)

40%

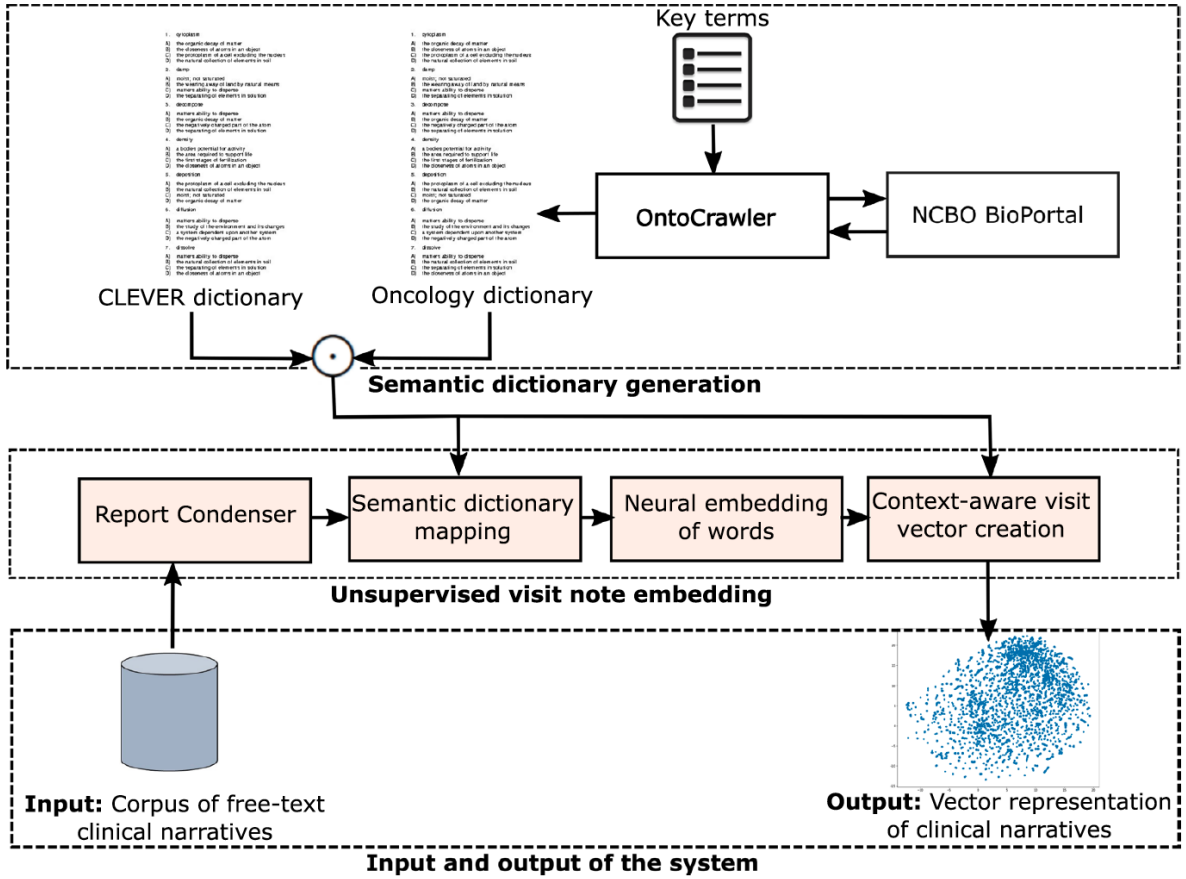
NLP Toolkit

- Extraction tool:
  - Query from SQL Server: **SQL, PyODBC**
- Formatting and cleaning:
  - **NLTK**: gold standard for Python
  - **Spacy**: more up-to-date functions, alternative to NLTK
- Data management:
  - **Pandas**
- Modeling:
  - **Scikit-Learn, Pytorch**
- Other useful resources:
  - **MIMIC-III:** massive EHR structured and unstructured data of critical care
  - **Huggingface**: open-source platform with pre-trained language models

Applications of EMR in Brain Mets Management

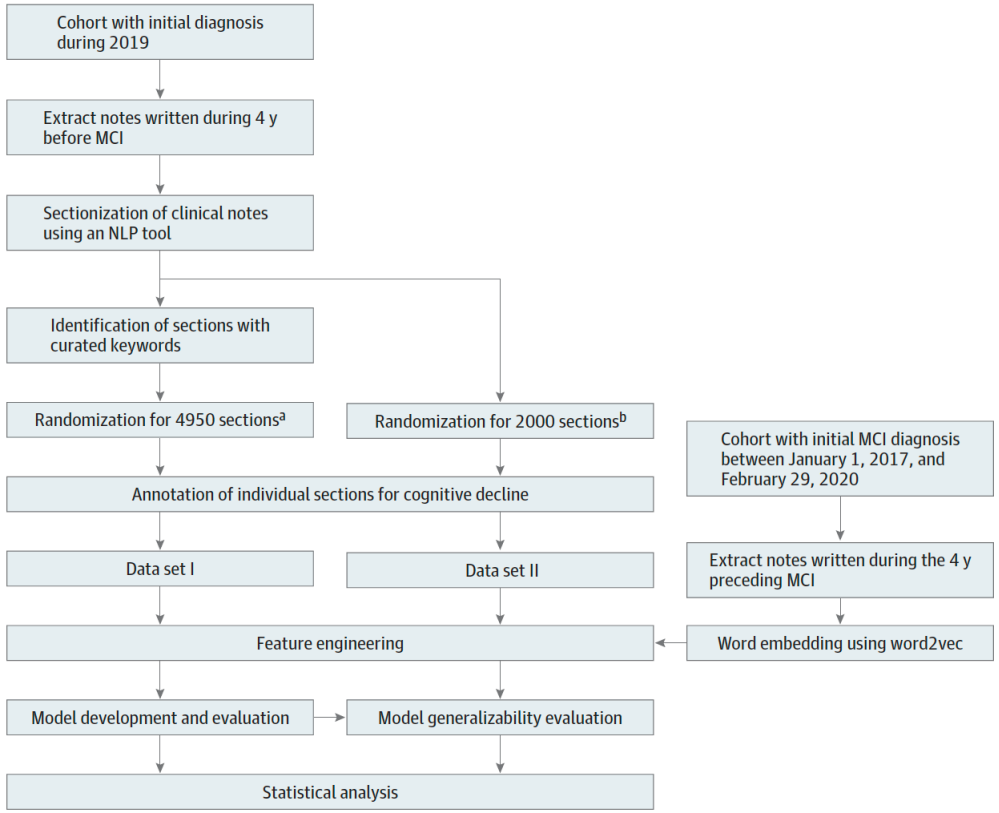
- Information retrieval
  - EMERSE: information retrieval from EHR free texts for patient cohort identification, data abstraction etc.
  - Sender et al.: quantification of brain mets in free-text radiology reports

- Survival prediction
  - Banerjee et al.: probabilistic prognostic estimate of short-term life expectancy by analyzing free-text clinical notes



Applications of EMR in Brain Mets Management

- Early detection of cognitive decline
  - Wang et al.: Detection of cognitive decline from unstructured clinical notes preceding MCI diagnosis



A. Attention to words contributing to the prediction of positive cases

Reason for referral Ms. Smith is a 78-year-old right-handed individual with sixteen years of formal education . She was referred for a neuropsychological evaluation to determine her current level of cognitive functioning as part of an ongoing effort to quantify any deficits that may be present in order to determine if the pattern of these deficits can provide insight into the process that induced them . She presents with concerns about her ability to remember things . She came to this appointment accompanied by her niece who helped to provide insight into the problems she has been having .
Objective measures : . A MOCA was completed during this consultation . Her overall score was a 23/30 , indicating potential mild cognitive impairment . Patient scored as follows : . Trail making exercise : 0/1 . Cube : 1/1 . Clock : 2/3 (hands/time incorrect) . Animal pictures : 3/3 . Number repetition : 2/2 . Letter recognition : 1/1 . Serial 7 : 2/3 . Sentence repetition : 2/2 . F-words : 1/1 . Abstraction : 2/2 . Delayed recall : 0/5 (2 words with category cue , 1 with multiple choice) . Orientation : 6/6 .
Date of onset Mr. Smith first noticed cognitive changes one year ago ; he opined that forgetfulness and word finding problems predated his stroke in February 2019 .

RadOnc Examples: MEDomics Consortium

nature cancer

ARTICLES

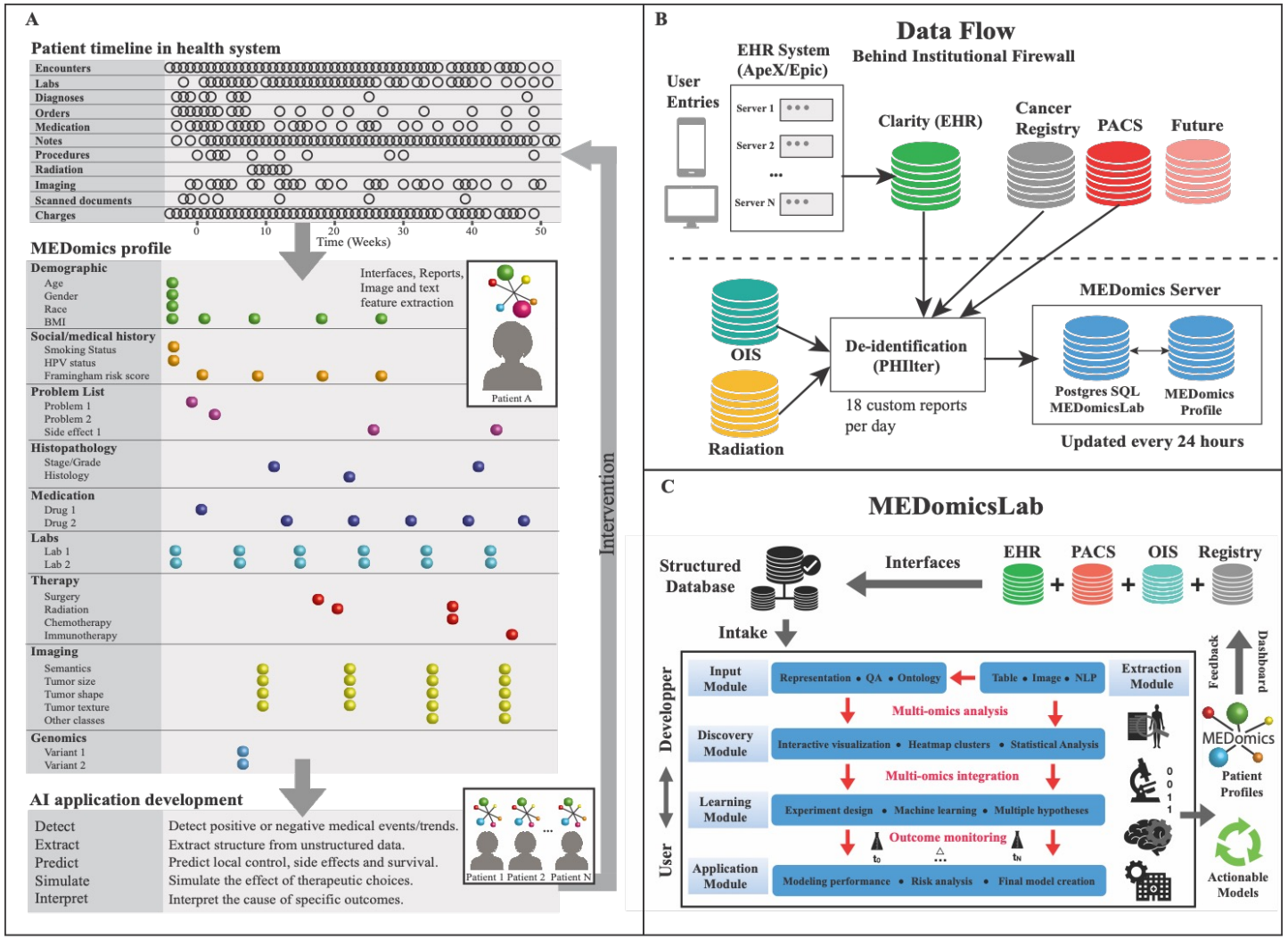
https://doi.org/10.1038/s43018-021-00236-2

Check for updates

An artificial intelligence framework integrating longitudinal electronic health records with real-world data enables continuous pan-cancer prognostication

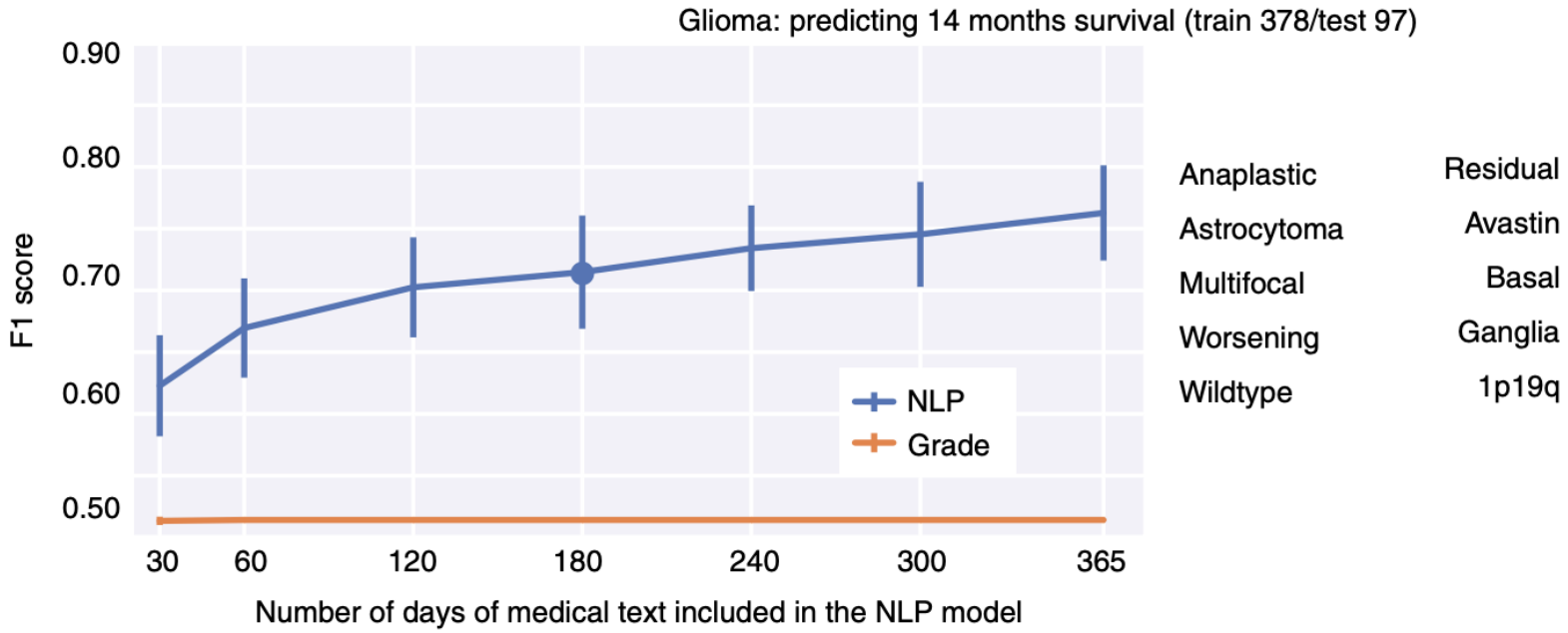
Olivier Morin[1✉], Martin Vallières[1,2,3], Steve Braunstein[1], Jorge Barrios Ginart[1], Taman Upadhaya[1], Henry C. Woodruff[4,5], Alex Zwanenburg[6,7,8,9,10], Avishek Chatterjee[2,4,5], Javier E. Villanueva-Meyer[11], Gilmer Valdes[1,12], William Chen[1], Julian C. Hong[1,13], Sue S. Yom[1], Timothy D. Solberg[1], Steffen Löck[6], Jan Seuntjens[2], Catherine Park[1] and Philippe Lambin[4,5]

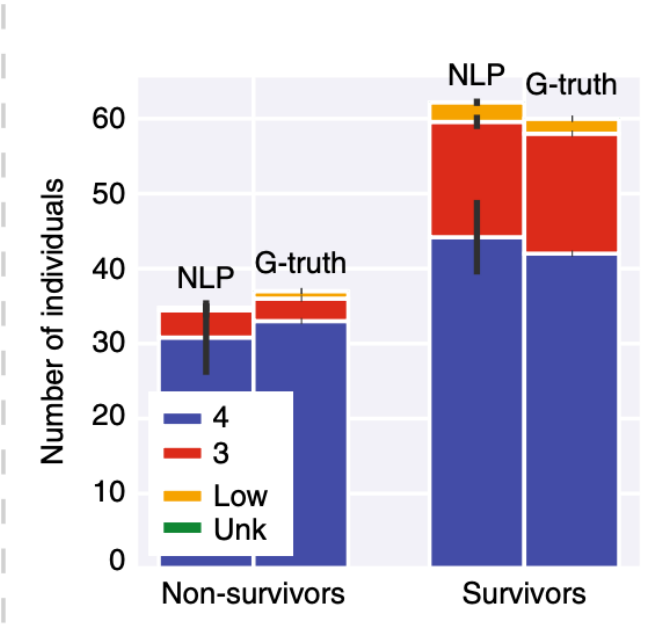
Despite widespread adoption of electronic health records (EHRs), most hospitals are not ready to implement data science research in the clinical pipelines. Here, we develop MEDomics, a continuously learning infrastructure through which multi-modal health data are systematically organized and data quality is assessed with the goal of applying artificial intelligence for individual prognosis. Using this framework, currently composed of thousands of individuals with cancer and millions of data points over a decade of data recording, we demonstrate prognostic utility of this framework in oncology. As proof of concept, we report an analysis using this infrastructure, which identified the Framingham risk score to be robustly associated with mortality among individuals with early-stage and advanced-stage cancer, a potentially actionable finding from a real-world cohort of individuals with cancer. Finally, we show how natural language processing (NLP) of medical notes could be used to continuously update estimates of prognosis as a given individual's disease course unfolds.



UCSF

Application: Processing unstructured EMR notes to reveal disease pattern and propel outcome prediction



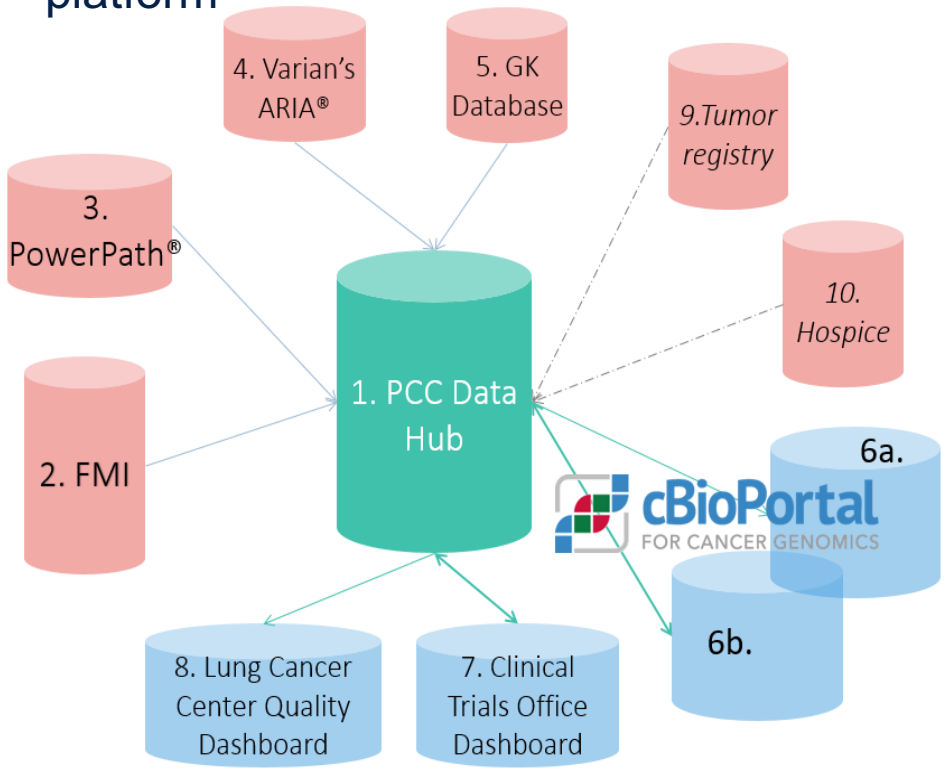


Morin, O. et al., 2021. Nature Cancer, 2(7), pp.709-722.

UCSF

RadOnc Examples: Data Hub at NYU Perlmutter Cancer Center

- Current and developing assets within the Data Hub. Assets are collocated on the Hadoop Data Lake platform



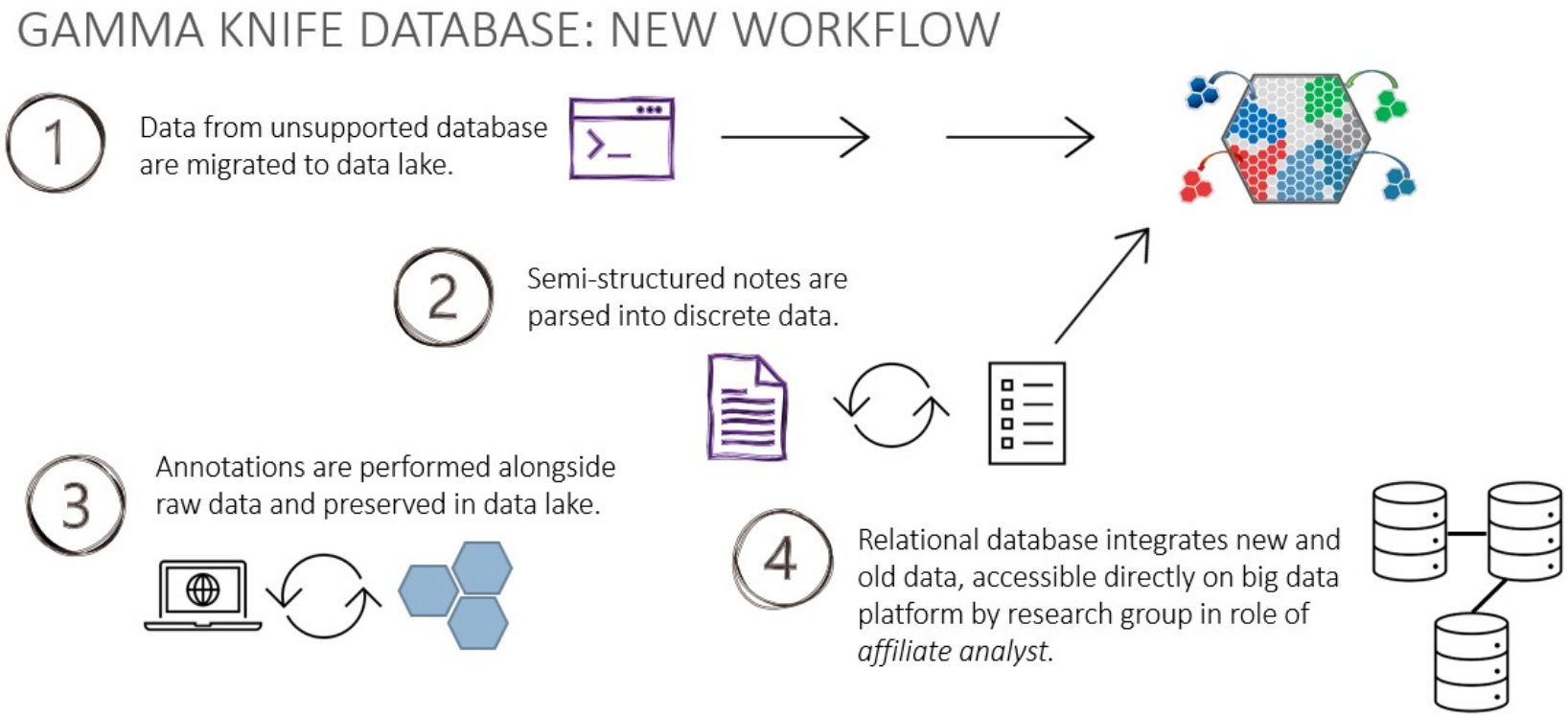
The Data Hub is a shared resource providing centralized, expert **data curation**, **storage**, **access**, and **analytic tools** to all Perlmutter Cancer Center (PCC) investigators at NYU Langone Health (NYULH). The Data Hub is co-directed by Dr. Megan Winner, MD MS FACS, Assistant Professor of Surgery, and Dr. Stephen Johnson PhD, Professor in the Department of Population Health, Director of the Clinical Research Informatics and Faculty Director of Datacore.

Slides in courtesy of Ken Bernstein @ NYU

1. **Gamma Knife prospective registry**
2. **Base tables derived from EPIC (a) structured data is translated into plane language (b) clinical notes are parsed into lines that are searchable with NLP tools**
3. **Third party tumor sequencing data for research**
4. **Legacy pathology data**
5. **ARIA/Eclipse Radiation treatment planning software**
6. **Death dates from Social Security Administration**
7. **Two instances of cBioPortal informing (a) clinical reporting and (b) research cohort exploration linked with biospecimens**
8. **Dashboard of data integrated from (2) and (3) to power clinical trials enrollment**
9. **Dashboard of data integrated from (2) and (3) to inform LCC quality metrics**
10. **Dashboard of data integrated from (1) and (2) for SRS outcome analysis**

NYU Langone Health

- The largest hurdle to integrating a prospective radiosurgery outcomes registry to a big data platform is patient matching:
  - MRN change with software, merging of hospital systems, typos and human error
  - Names change with marriage, misspellings, abbreviations, nicknames, preferred names



Slides in courtesy of Ken Bernstein @ NYU

Acknowledgments

- Medomics NLP Consortium


Olivier Morin, PhD


Jorge Barrios, PhD


Will C Chen, MD


Yannet Interian, PhD

- Kenneth Bernstein @ NYU

- Robert Thombley and Dana Ludwig's talks @ UCSF NLP group

UCSF

References

Carrell, D.S. et al., 2017. Journal of the American Medical Informatics Association, 24(5), pp.986-991.

Hanauer, D.A. et al., 2020. JCO clinical cancer informatics, 4, 454-463.

Senders, J.T. et al., 2019. JCO Clinical Cancer Informatics, 3, pp.1-9.

Banerjee, I. et al., 2018. Scientific reports, 8(1), pp.1-12.

Wang, L. et al., 2021. JAMA network open, 4(11), pp.e2135174-e2135174.

Ling, A.Y. et al., 2019. JAMIA open, 2(4), pp.528-537.

Zeng, Z. et al., 2019. Journal of healthcare informatics research, 3(3), pp.283-299.

Morin, O. et al., 2021. Nature Cancer, 2(7), pp.709-722.

Huang, K. et al., 2019. arXiv preprint arXiv:1904.05342.

Thank you!

University of California
San Francisco
advancing health worldwide™

Hui Lin, PhD
Department of Radiation Oncology
UCSF
hui.lin@ucsf.edu