EMORY UNIVERSITY SCHOOL OF MEDICINE

Department of Biomedical Informatics

# BMI 500: https://tinyurl.com/bmi500

# Introduction to Biomedical Informatics

## 6. Text Representations, Comparisons and Knowledge Sources

28th Sept, 2022

Abeed Sarker

Department of Biomedical Informatics, Emory University, Atlanta, GA USA

# Recap

- What is Natural Language Processing (NLP)?
- Why is NLP important?
- What are some of the basic challenges for NLP?
- What has NLP accomplished so far, particularly in biomedical informatics?
- What is the future of NLP?

# Expectations: Deliverables

- Participation in class
- Exploring and comparing texts using NLP
- End-to-end (*full stack*) NLP pipelines

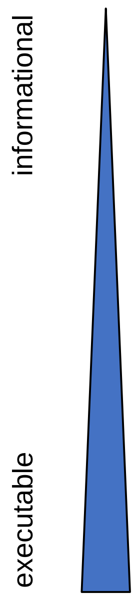
# Overview

- Part 1
  - Data and knowledge storage
    - Vocabularies
    - Ontologies
- Part 2
  - NLP basics
- Part 3
  - End-to-end BioNLP systems

# Part 1

# Encapsulating medical knowledge

# Different Ways of Storing Knowledge




- *Data dictionary*
  A set of terms and their definitions
- *Controlled Vocabulary*
  A terminology designed for a particular information retrieval application
- *Ontology*
  A controlled vocabulary expressed in an ontology representation language

*Controlled* means adhering to local conventions or to terms set by an external standards body.

*This slide is from the 2018 BMI500 lecture prepared by Andrew Post

# Unified Medical Language System

- https://www.nlm.nih.gov/research/umls/
- Repository of biomedical vocabularies
- Web site for browsing and searching ontologies (sign-up required)
- Can download all vocabularies and mappings as CSV files (some vocabularies require special licensing)
- Every vocabulary term maps into a UMLS concept called a concept unique identifier (CUI)
- Represents mappings between terms (through the CUI)
- Contains rich store of synonyms for every concept that can improve the accuracy of natural language processing algorithms

*This slide is from the 2018 BMI500 lecture prepared by Andrew Post

# BioPortal

- https://bioportal.bioontology.org
- Repository of biomedical ontologies
- Web site for browsing and searching ontologies
- Web services APIs for accessing ontologies
- Supports downloading ontologies in OWL and other formats
- Widgets for selecting concepts in web applications

*This slide is from the 2018 BMI500 lecture prepared by Andrew Post

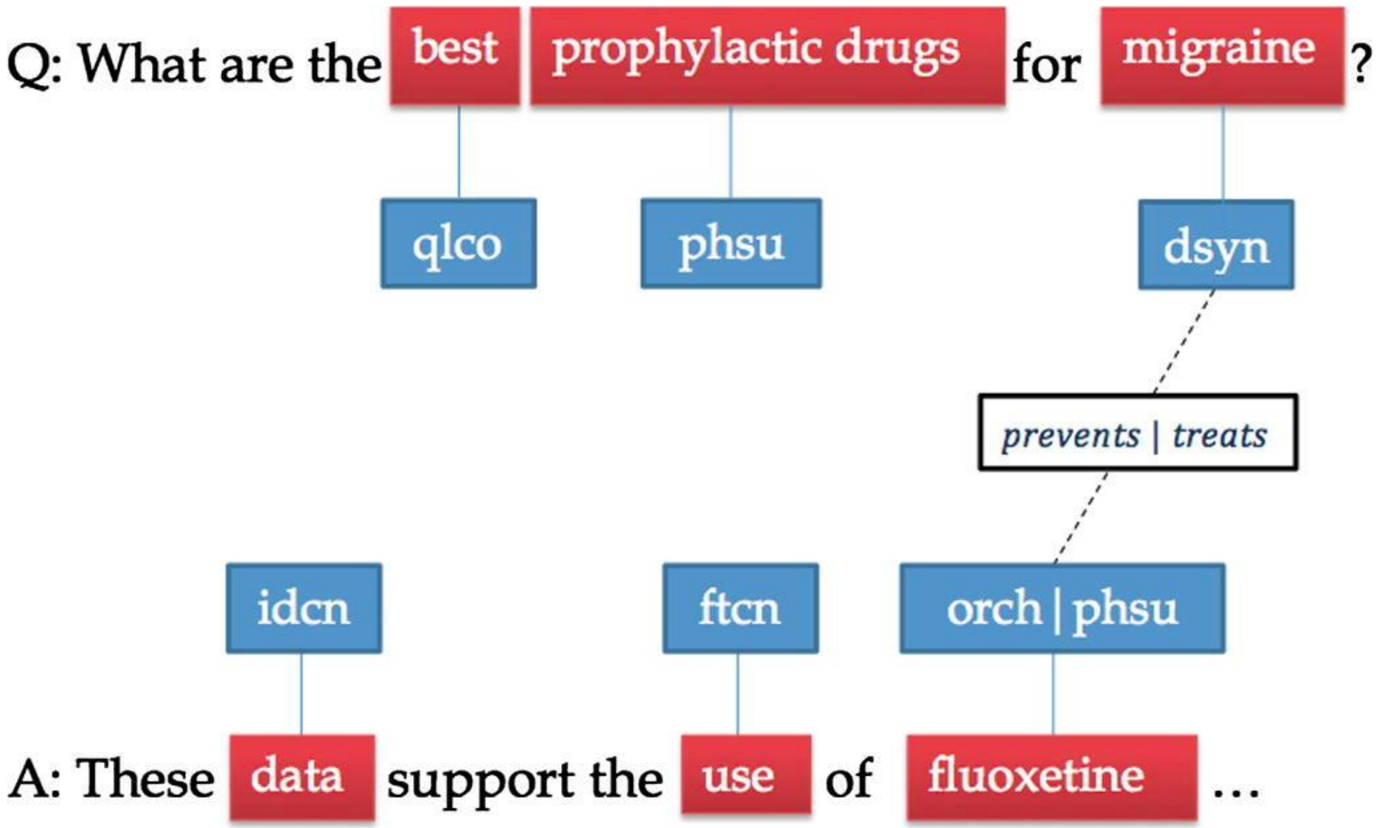
# PubMed

MeSH Unique ID: D011471
Entry Terms:

- Prostate Neoplasms
- Neoplasms, Prostate
- Neoplasm, Prostate
- Prostate Neoplasm
- Neoplasms, Prostatic
- Neoplasm, Prostatic
- Prostatic Neoplasm
- Prostate Cancer
- Cancer, Prostate
- Cancers, Prostate
- Prostate Cancers
- Cancer of the Prostate
- Prostatic Cancer
- Cancer, Prostatic
- Cancers, Prostatic
- Prostatic Cancers
- Cancer of Prostate

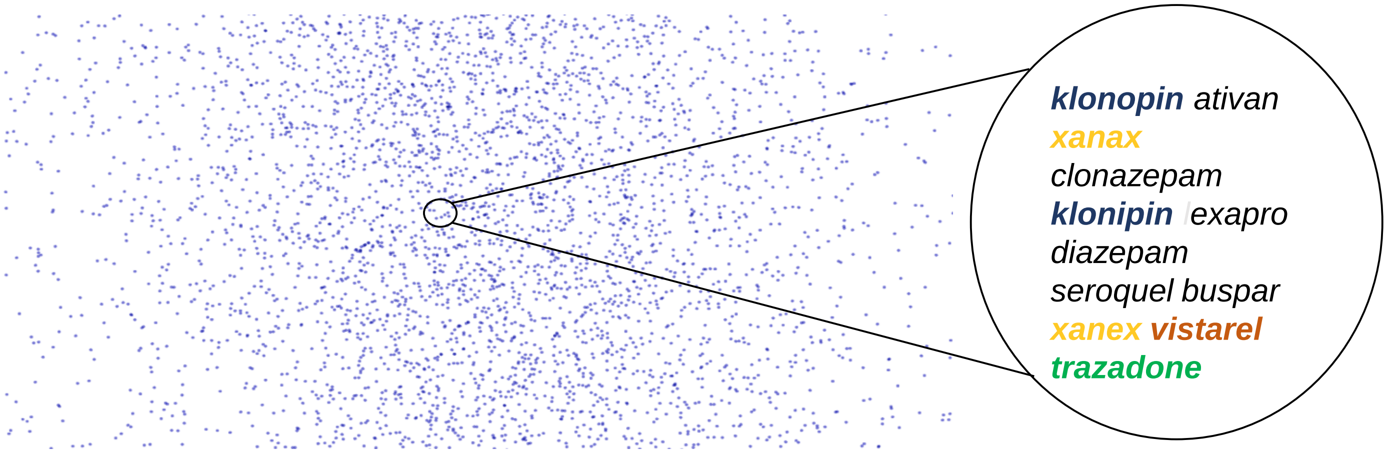

*This slide is from the 2018 BMI500 lecture prepared by Andrew Post

# Using knowledge resources in medical NLP



# Knowledge capture via text representations

- N-gram-based text representations do not capture meanings of terms or phrases
  - Sparse vectors
- Word/phrase-embeddings (word2vec, GLoVe) capture meanings of short text segments
  - Dense vectors
  - They do not capture contextual variations in meanings



# Part 2

# NLP Basics

# NLP Practical

- text representation
  - n-grams
- tagging

# Text representation – n-grams

- Contiguous sequence of n items from a given sequence
- Typically words or characters
- N-gram of size 1: unigram
  - Size 2: bigrams, Size 3: trigrams and so on
- Most vector representations that we have looked at are generated from n-grams in real life

# Text representation – n-grams

- *`Contiguous sequence of n items from a given sequence`*
- Unigrams `['Contiguous', 'sequence', 'of', 'n', 'items', 'from', 'a', 'given', 'sequence']`
- Bigrams `[('Contiguous', 'sequence'), ('sequence', 'of'), ('of', 'n'), ('n', 'items'), ('items', 'from'), ('from', 'a'), ('a', 'given'), ('given', 'sequence')]`
- Trigrams `[('Contiguous', 'sequence', 'of'), ('sequence', 'of', 'n'), ('of', 'n', 'items'), ('n', 'items', 'from'), ('items', 'from', 'a'), ('from', 'a', 'given'), ('a', 'given', 'sequence')]`

# N-grams in nltk

- N-grams may be used in the same way as individual tokens
  - Document comparisons are likely to be more reliable when n-grams are used rather than just tokens (unigrams)

- Vectorization techniques remain the same

- N-gram vectors are sparser

- Despite the sparsity, these n-grams may capture crucial sequence information in text
  - ‘I am new in new york’
  - Bag of words representation does not capture if the person is new in ‘New York’ or ‘York’

# Text tagging and representations

- More information on top of n-grams can be added via NLP
- POS tagging is a common task
- The process of marking up a word in a text as corresponding to a  particular part of speech based on
  - Definition
  - Context
- The first step in semantic analysis of text
  - *e.g., 'The checkout person bags the products'* *vs. 'There were too few bags at  the checkout counter'*
- For text in the medical domain, there are typically many ways tagging can be done
  - *e.g.*, terms representing diseases, adverse reactions, treatments etc.

# Practical problems with tagging of clinical texts

- Taggers *(aka*. Entity recognizers, named entity recognizers) need to be trained using the type of text on which they are to be applied
- Clinical texts such as from EHRs are typically not publicly available
  - So aren't their annotations
- As a result, research groups have to annotate their data in-house to create training data
  - … and it's not an easy annotation task

# Comparing documents—vector representations of texts

- The simplest mechanism to compare documents is to compare the overlap in word types
  - Documents discussing similar topics are likely to have high overlap in word types
- Jaccard similarity is a common measure
- *Jaccard_similarity =* $(s1 \cap s2)/(s1 \cup s2)$, where s1 and s2 are the sets of word types in documents **s1** and **s2**

# Example

- s1 = [**'this'**,**'is'**,**'an'**, **'example'**]
- s2 = [**'this'**, **'is'**, **'a'**, **'separate'**, **'sentence'**]
- $s1 \cap s2$ = ['this', 'is'] (length = 2)
- $s1 \cup s2$ = ['a', 'sentence', 'this', 'is', 'separate', 'an', 'example'] (length =7)
- Jaccard_similarity = 2/7 = 0.286

# Vector representation of text

- Most similarity and other techniques rely on vector-based representations of texts
- First step in creating a vector-based representation:
- Creating a vocabulary
- A vocabulary consists of all the word types in a document set or a corpus
- A corpus is a set of texts that are used together for some useful task

# Vector representation of text

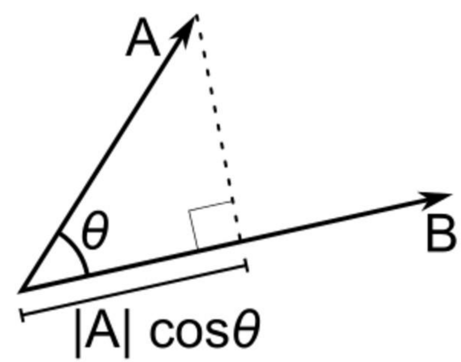
- The simplest vector representation uses 0s and 1s to indicate the absence or  presence of a term in a vocabulary
- s1 = ['this','is','an', 'example']
- s2 = ['this', 'is', 'a', 'separate', 'sentence']
- Vocabulary = 'a', 'sentence', 'this', 'is', 'separate', 'an', 'example’
- The vocabulary has size = 7
- s1 -> [0,0,1,1,0,1,1]
- s2 -> [1,1,1,1,1,0,0]

# Sparsity of vectors

- Such vectors are sparse in nature
- In a real-life task, only a small number of columns will be 1
- Imagine the representations for the same sentences when the vocabulary size is 10,000
- The sparsity can be problematic – very recent advances in NLP have led to dense vector representations
- This model is also called: bag-of-words model
- What other information is not preserved in these vectors?

# Cosine similarity

- Cosine of the angle between two  vectors
- Cosine similarity generates a  value that shows how related two  document vectors are
- cos(0) = 1
- cos(1) = 0
- Very popular
- Scales to vectors of any number of dimensions
- Can be applied to other document vector repre
- Works for sparse vectors
- Fast



# Other document vector representations

- Term frequency
- TF-IDF (term frequency-inverse document frequency)
- Intended to reflect how important a word is for a document in a corpus
- A term occurring frequently in a document will have high TF value
- A term occurring frequently across the corpus will have high DF value
- The product of TF with the inverse of DF (IDF) gives the TF-IDF value
- Very powerful representation; used by many text processing systems

# Dense vector representations and current SOT

- Dense vector models capture complex information
- BERT-based models have revolutionized long text vector representations and have improved performances on many tasks… but…
  - Medical text is still very complex
  - Current dense models cannot capture much of the complex associations

- For some tasks, similar performances can be achieved for domain-independent and domain-specific texts

# NLP Evaluations (outline only)

- Intrinsic *vs.* extrinsic evaluations

- Evaluation metrics
  - Accuracy (classification, extraction, normalization…)
  - Precision, recall, $F_1$-score

- Complex evaluations
  - e.g., ROUGE (summarization)

- Confidence intervals

- Statistical tests

- Do evaluation metrics actually make sense?
  - Inter-annotator agreements

- Possible reading (Resnik & Lin, 2013): https://www.cs.colorado.edu/~jbg/teaching/CMSC_773_2012/reading/evaluation.pdf

# Part 3

# End-to-end BioNLP systems

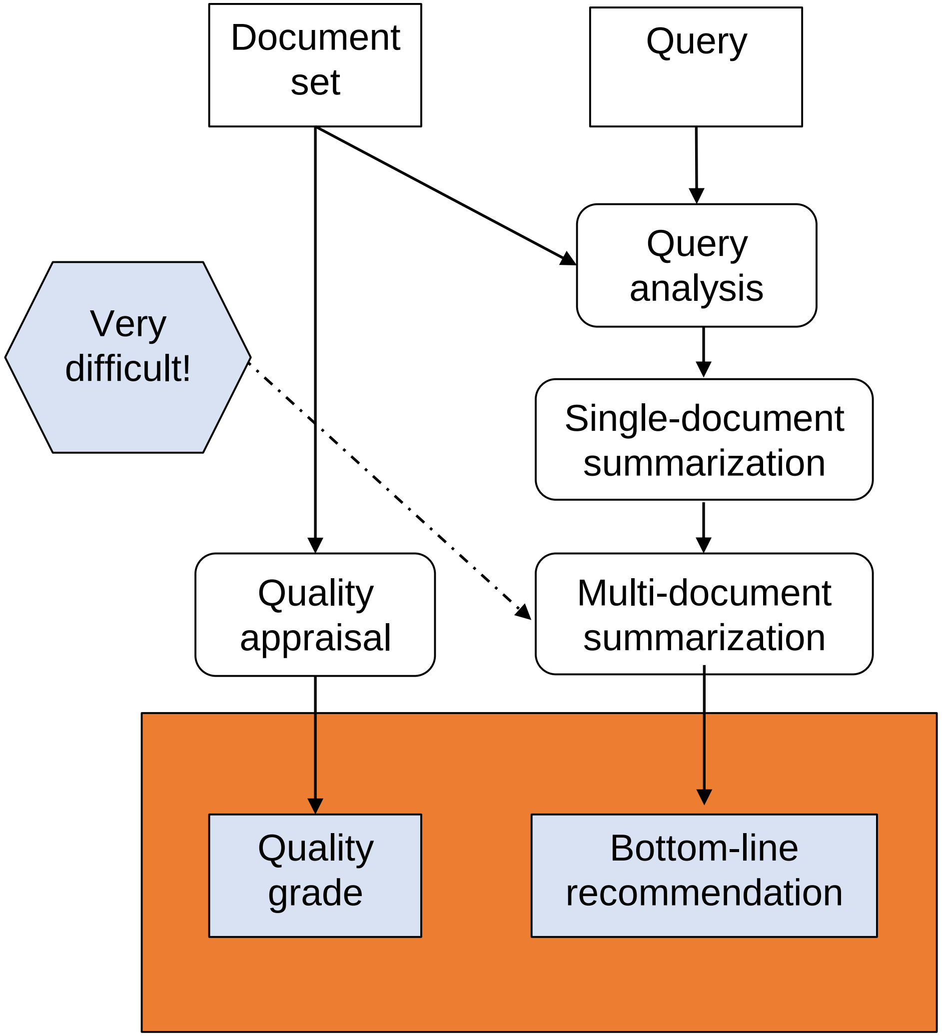
# End-to-end architecture—text summarization

**Query:** Are there big differences in beta-blockers in treating essential hypertension?

**Automatic extractive summary:**
Because the pathophysiology of hypertension differs in older and younger patients, we designed this meta-analysis to clarify the efficacy of beta-blockers in different age groups.
In placebo-controlled trials, beta-blockers reduced major cardiovascular outcomes in younger patients (risk ratio [RR] 0.86, 95% confidence interval [CI] 0.74-0.99, based on 794 events in 19 414 patients) but not in older patients (RR 0.89, 95% CI 0.75-1.05, based on 1115 events in 8019 patients).
Beta-blockers should not be considered first-line therapy for older hypertensive patients without another indication for these agents; however, in younger patients beta-blockers are associated with a significant reduction in cardiovascular morbidity and mortality.

(Quality of evidence: **A**)

**PMID:** 16754904



# Domain comparison

**Domain independent text**

**Question:** *Who wrote the Foundation Series?*

NLP system tasks:

- Quite straightforward query analysis
- '*who*' indicates the question is looking for the name of a person
- Any public document on the topic can be used to easily extract the answer
- *Factoid question*
- Some summarization tasks are relatively easy (*e.g.*, news summarization)
- Can get more complex though…

**Medical text**

**Question:** Are there big differences in beta-blockers in treating essential hypertension?

NLP system:

- Needs to know what a beta-blocker is. Medical articles often refer to the specific medication/intervention.
- Also, effectiveness of drug can vary based on patient attributes (*e.g.*, age, gender etc.)
- "*big*" is not very well defined.
- Are there different forms of hypertension?
- NLP methods, without domain knowledge, will not be useful.

# End-to-end architecture 1—text summarization for evidence-based medicine

- Query analysis
  - Identifying the medical subdomain of the question (*e.g.*, treatment, diagnosis, prognosis etc.)
- Single-document summarization
  - Identifying document texts that are relevant to the query
- Multi-document summarization
  - Combining evidence from multiple documents
- Assessing the quality of medical evidence
  - Randomized controlled trials *vs.* case report

## Our publications on the topic (selected)


**Single-document summarization:**

1. Sarker A, Mollá D, Paris C**. Query-oriented evidence extraction to support evidence-based medicine practice**. J Biomed Inform. 2016;59:169-184. doi:10.1016/j.jbi.2015.11.010

2. Sarker A, Mollá D, Paris C. **Extractive evidence based medicine summarisation based on sentence-specific statistics**. 2012 25th IEEE International Symposium on Computer-Based Medical Systems (CBMS), Rome, 2012, pp. 1-4, doi: 10.1109/CBMS.2012.6266373.

**Multi-document summarization (information synthesis):**

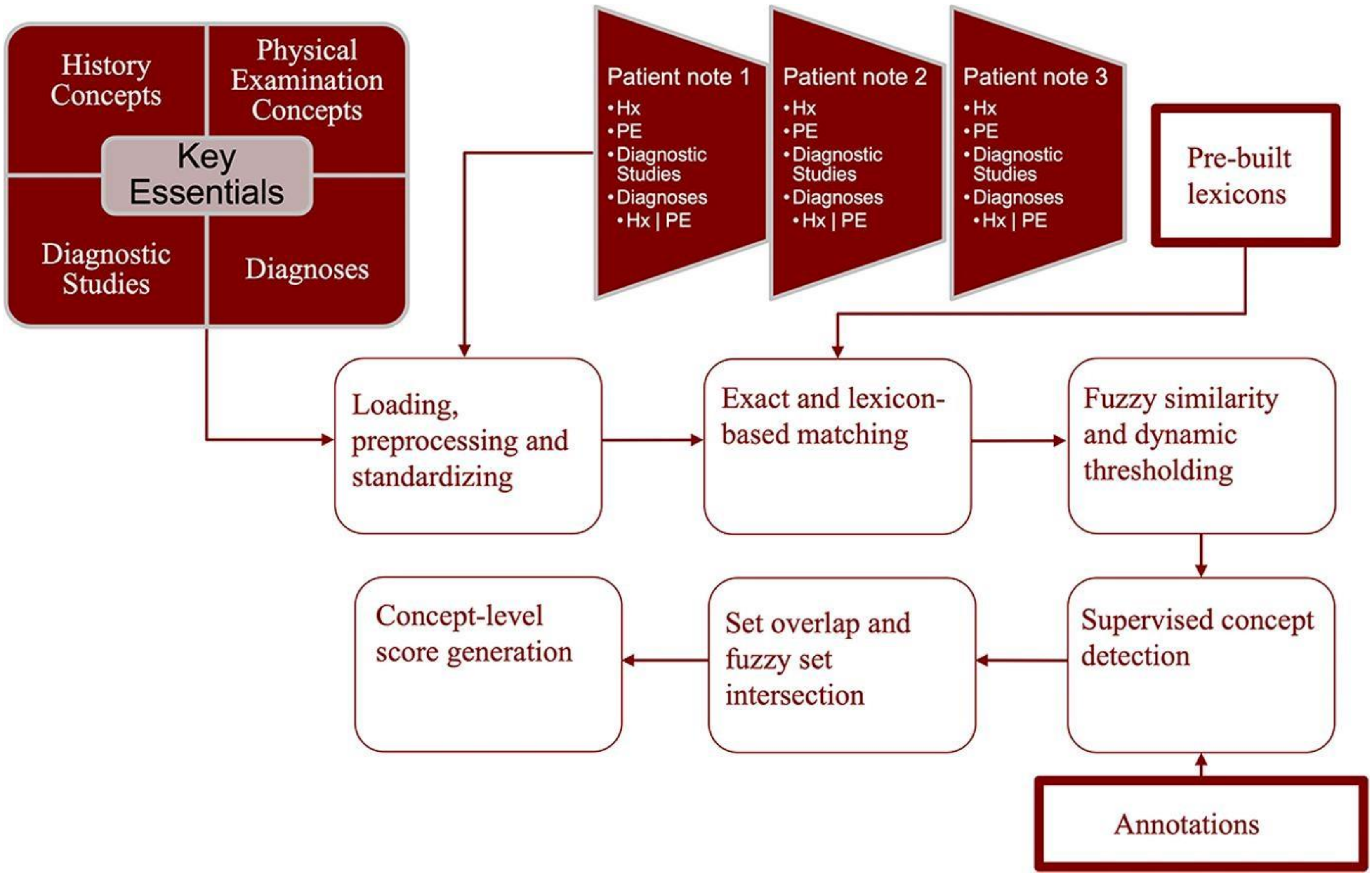
3. Sarker A, Mollla-Aliod D, Paris C. **Automatic Prediction of Evidence-based Recommendations via Sentence-level Polarity Classification**. International Joint Conference on Natural Language Processing, pages 712–718, Nagoya, Japan, 14-18 October 2013.

**Quality assessment:**

4. Sarker A, Mollá D, Paris C. **Automatic evidence quality prediction to support evidence-based decision making**. *Artif Intell Med*. 2015;64(2):89-103. doi:10.1016/j.artmed.2015.04.001

# End-to-end architecture 2—detecting specialized clinical concepts from EHRs

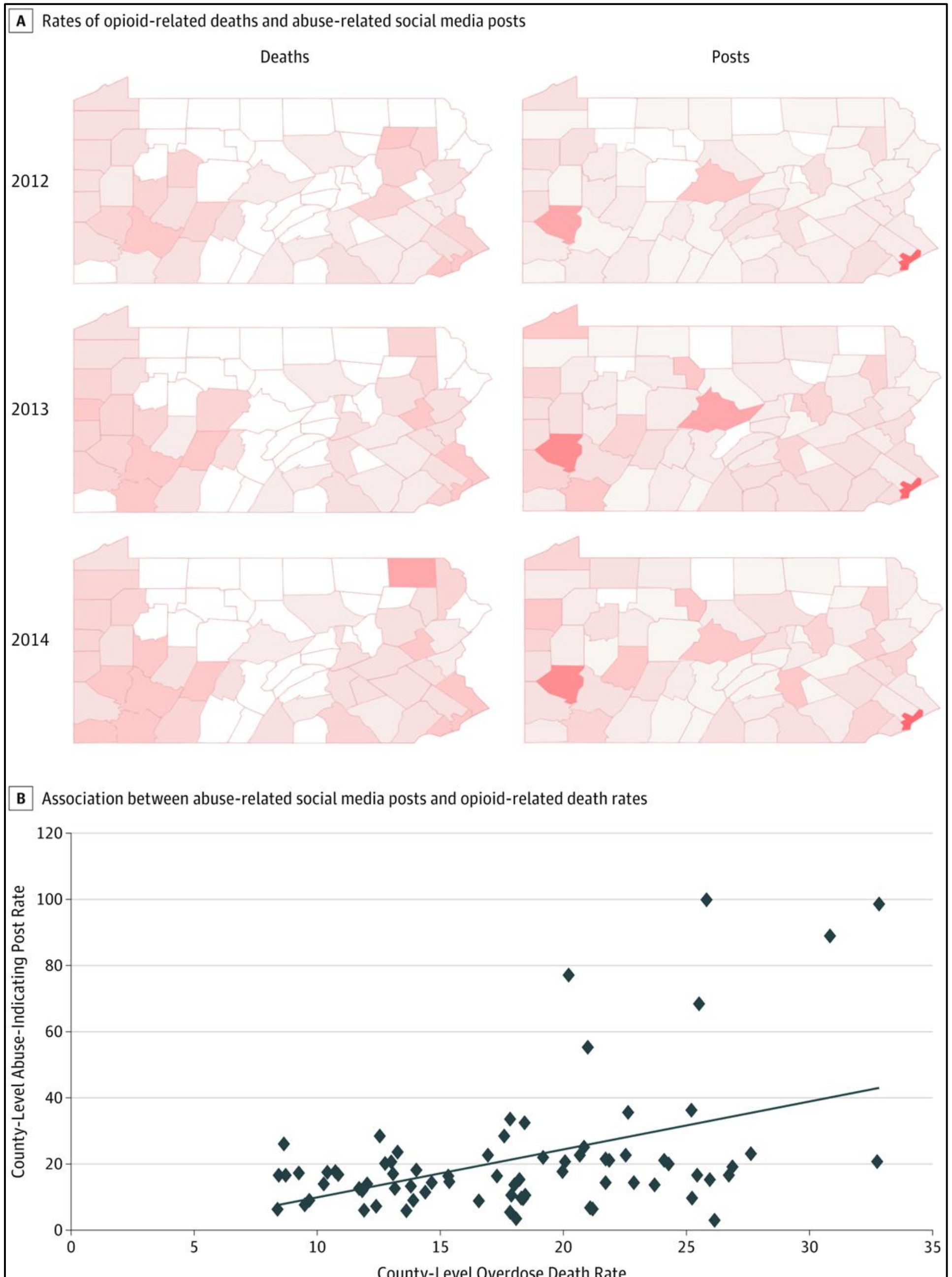
- Task—specify selected ‘*ad-hoc*’ concepts from EHRs.
- How it happens in real life:
  - Doctor/researcher wants to study a specific information about patients.
  - Information includes different types of patient-level information (*e.g.*, age, gender), health-related information *(e.g*., diagnostic results, history of health events, medications etc.), and other targeted information *(e.g*., travel, contact with infected people etc.)
- CS/machine learning/NLP tasks typically focus on just 1 of these (*e.g.*, detecting disease or adverse reaction from free text)
- Real life problem has many constraints

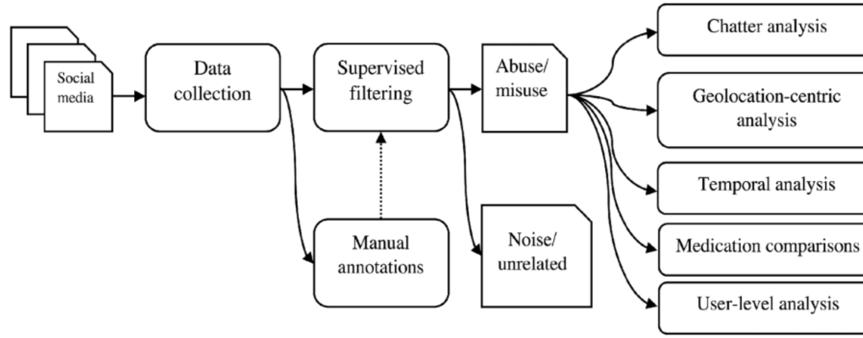



Sarker A *et al*. An interpretable natural language processing system for written medical examination assessment. J Biomed Inform. Volume 98, 2019, 103268, ISSN 1532-0464, https://doi.org/10.1016/j.jbi.2019.103268.

# End-to-end architecture 3—social media mining for toxicovigilance

- Tasks
  - Identify discussions about drugs/medications on social media
  - Distinguish between medical use, nonmedical use, and other types of information
  - Analyze the contents of the chatter
  - Identify geographic patterns and temporal patterns




Sarker A, DeRoos A, Perrone J. Mining social media for prescription medication abuse monitoring: a review and proposal for a data-centric framework. J Am Med Inform Assoc. 2020;27(2):315-329. doi:10.1093/jamia/ocz162.

Sarker A, Gonzalez-Hernandez G, Ruan Y, Perrone J. Machine Learning and Natural Language Processing for Geolocation-Centric Monitoring and Characterization of Opioid-Related Social Media Chatter. JAMA Netw Open. 2019;2(11):e1914672. Published 2019 Nov 1. doi:10.1001/jamanetworkopen.2019.14672.

# NLP lab work (week 6)

- This week's lab work will involve
  - Text representation
  - POS tagging
  - + other tasks..
- Please find the Lab tasks here:
  - https://drive.google.com/file/d/11P8K3A4j2nHr4Yj_x7pibI8U6dP40Rq5/view?usp=sharing
  - Files are available here: https://drive.google.com/file/d/1pxF3KsULkKR9UvpByw__ezptL-j5-UcL/view?usp=sharing
- Solutions will be posted after submission

# RECAP

NLP is an important field within the health research space as most medical knowledge is encapsulated in free text form

NLP of biomedical texts is more challenging than domain-independent text

The primary outcomes (expected) from the lectures outlining NLP are:

- An understanding of the relevance of NLP
- Basic NLP including preprocessing, searching/matching, text representation and evaluation
- An understanding of the complexities associated with building end-to-end (full stack) NLP systems for health-related texts