Improve Chart Review Efficiency using Natural Language Processing and Machine Learning

Tianrun Cai, M.D.

Associate Bioinformatician
Brigham and Women's Hospital
Instructor in Medicine
Harvard Medical School

VERITY RESEARCH

BRIGHAM HEALTH
BRIGHAM AND WOMEN'S HOSPITAL

HARVARD MEDICAL SCHOOL TEACHING HOSPITAL

Background

- Electronic Medical Record (EMR)
  - Adopted by most of hospitals in US
  - Large amount of EMR data available for research
  - More and more notes for each patient

Background

- Big data enables large-scale research
  - Phenome Wide Association Studies (PWAS)
    - Definition: the association between single-nucleotide polymorphisms or other types of DNA variants is tested across a large number of different phenotypes.

Example: IL6R variant identifies a drug target for cardiovascular disease and inflammation[1]

- **342k** US veterans; UK Biobank, **408K**
- **1,866** phenotype groups
- Manual chart review for **20 phenotypes**

[1]. ***Tianxi Cai,*** *et al. "Association of interleukin 6 receptor variant with cardiovascular disease effects of interleukin 6 receptor blocking therapy: a phenome-wide association study." JAMA cardiology 3, no. 9 (2018): 849-857*

Background

- One of **major bottlenecks** of analyzing EMR data– precise phenotypes
- Phenotyping algorithms
  - Rule-based approaches
  - Machine-learning approaches (supervised, unsupervised)
- **Manual chart review** is needed to obtain gold standard labels for either algorithm training and/or algorithm validation

Background

Challenge of chart review: Too many notes

- ❑ Large number of clinic notes per patient
  - VA hospitals: Range: 7-5037
    average: 262 notes/patient,
    10% patients have >1000 notes
  - Partners Biobank : Range: 1-3154,
    average: 170 notes/patient
- ❑ Reviewing all notes is ideal but infeasible
  - Reviewing notes selectively - string search

Background

- **Example: Chart review for phenotype Rheumatoid Arthritis**
  - Search terms for RA to get notes: **morning stiffness, joint pain, arthralgia, rheumatoid arthritis**, or **RA**
  - Miss notes only contain **painful joints**, **knee pain**, **ankle pain** etc.
  - **How we can find a systematic way to get notes related to a phenotype of interest**

Note ranking algorithm

Note ranking algorithm – Systematically ranking notes by relevance

Which notes are more informative ?

- Notes with many mentions of Rheumatoid Arthritis (RA) related **concepts***(RA, morning stiffness…) tend to be informative.

Background

“Concept” and “CUI”

1. A concept – meaning of a term, a concept stands for a group of synonyms
   - Concept “**Joint Pain**” : joint pains, aching joint, arthralgia, etc.
2. CUI – Concept Unique Identifier assigned in Unified Medical Language System (UMLS) for each medical concept
   - Joint pain:**C0003862**

Background

- Example notes with different number of **RA related concepts**

…
HISTORY: Multiple sclerosis, gastroesophageal reflux disease, **joint pains** in her knees, overweight, **inflammatory Polyarthritis** with mildly increased **CRP**.
…
Reason for Visit
Joint Pain
Reason for Visit History
Diagnoses
Inflammatory arthritis - Primary M19.90
…

Note 1

……
I started having **joint pains in both wrists** 3 month ago.
…

One of my friends told me there is a test that can be done to see if you have **arthritis**.
……

Note 2

Background

How can we build a note ranking algorithm to understand this?

- Notes with many mentions of RA relevant concepts are informative to RA

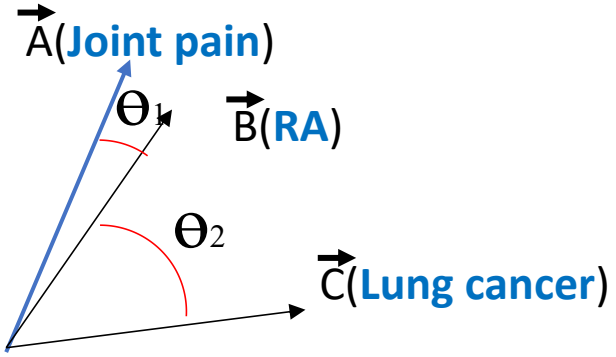
Key technologies

❑ CUI-to-Vec

- Each CUI →a semantic ***VEC***

arthritis,C0003864 → [0.273,0.371,-0.064,0.117,-0.21….]

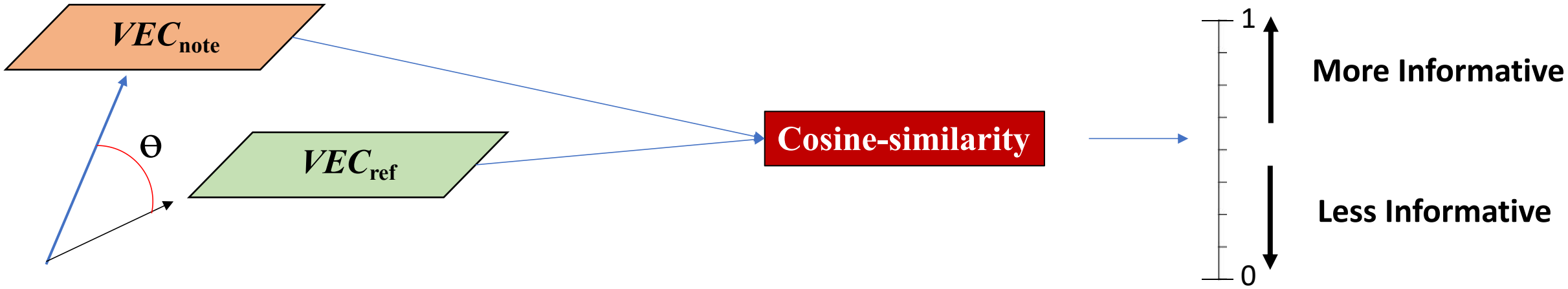
❑ Cosine similarity →Semantic similarity analysis



$$\text{Similarity}(\vec{A}, \vec{B}) = \cos(\theta_1)$$
$$\text{Similarity}(\vec{B}, \vec{C}) = \cos(\theta_2)$$
$$\text{Similarity}(\vec{A}, \vec{B}) > \text{Similarity}(\vec{B}, \vec{C})$$

Key ideas

- Summarize each note into a VEC_{note} with respect to RA content
- Compare the VEC_{note} to a VEC_{ref}
  - Summarize knowledge sources such as Wikipedia into a VEC_{ref}
  - Measure similarity between VEC_{note} and VEC_{ref}



Note Processing

- Get CUI count for each note using NLP software

inflammatory arthritis	C0003864	1
polyarthritis	C0162323	1
joint pain	C0003862	2
chronic pain	C0150055	1
CRP	C0006560	1

CUI count of Note 1

joint pain in both wrists	C2108718	1
arthritis	C0003864	1

CUI count of Note 2

Note Processing

- Get CUI count for the Wikipedia article related to the topic RA

Rheumatoid arthritis

From Wikipedia, the free encyclopedia

For juvenile rheumatoid arthritis, see juvenile idiopathic arthritis

Rheumatoid arthritis (**RA**) is a long-term autoimmune disorder that primarily affects joints.[1] It typically results in warm, swollen, and painful joints.[1] Pain and stiffness often worsen following rest.[1] Most commonly, the wrist and hands are involved, with the same joints typically involved on both sides of the body.[1] The disease may also affect other parts of the body.[1] This may result in a low red blood cell count, inflammation around the lungs, and inflammation around the heart.[1] Fever and low energy may also be present.[1] Often, symptoms come on gradually over weeks to months.[2]

While the cause of rheumatoid arthritis is not clear, it is believed to involve a combination of genetic and environmental factors.[1] The underlying mechanism involves the body's immune system attacking the joints.[1] This results in inflammation and thickening of the joint capsule.[1] It also affects the underlying bone and cartilage.[1] The diagnosis is made mostly on the basis of a person's signs and symptoms.[2] X-rays and laboratory testing may support a diagnosis or exclude other diseases with similar symptoms.[1] Other diseases that may present similarly include systemic lupus erythematosus, psoriatic arthritis, and fibromyalgia among others.[2]

Wiki article related to the disease RA

inflammatory arthritis	C0003864	15
antimalarial agent	C0003374	2
joint pain	C0003862	6
rheumatoid arthritis	C0003873	19
CRP	C0006560	4
…	…	…

CUI count of Wiki article

Summarize vectors for each note

- **Map CUIs to Vectors** and **summarize** the vectors for each note

Note 1

inflammatory arthritis	C0003864	1
polyarthritis	C0162323	1
joint pain	C0003862	2
chronic pain	C0150055	1
CRP	C0006560	1

Map →

VEC1	1
VEC2	1
VEC3	2
VEC4	1
VEC5	1

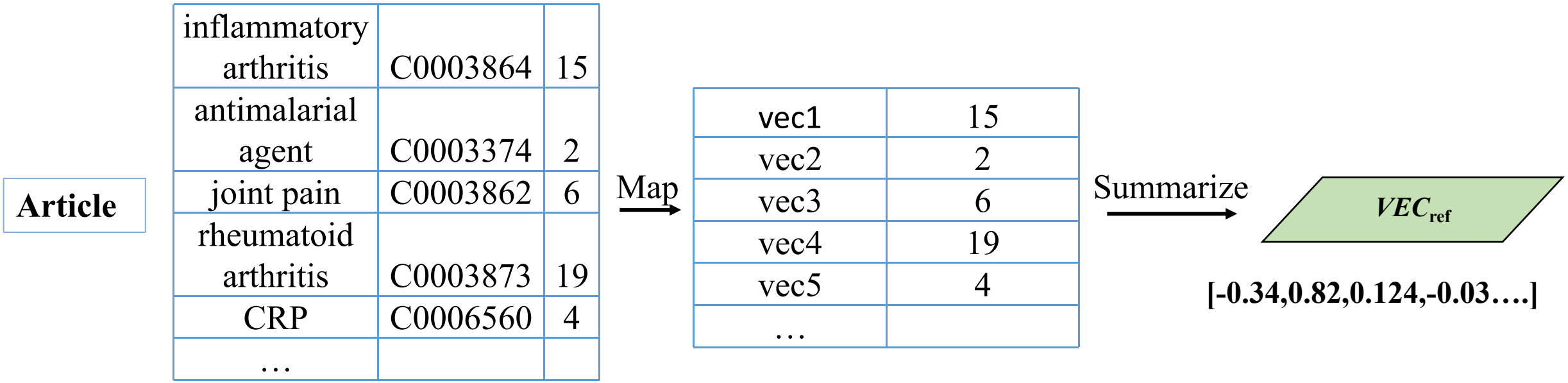
Summarize →

VEC_{note}

[0.03,-0.41,-0.147,0.218….]

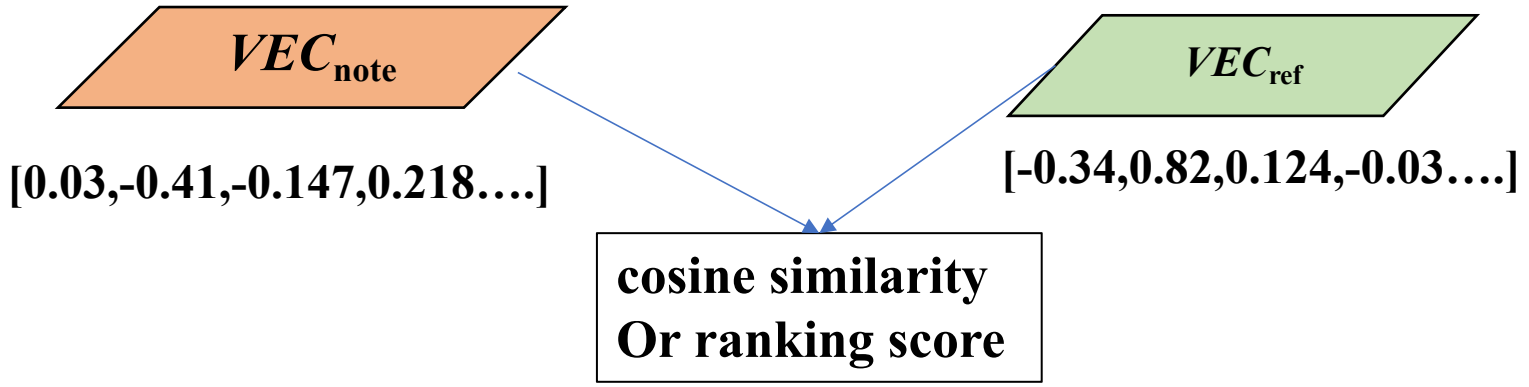
Summarize vectors for the article

- **Map CUIs to Vectors** for Wikipedia article and **Summarize** the Vectors into VEC_{ref}



Obtaining Ranking Scores

- Perform **cosine similarity** analysis

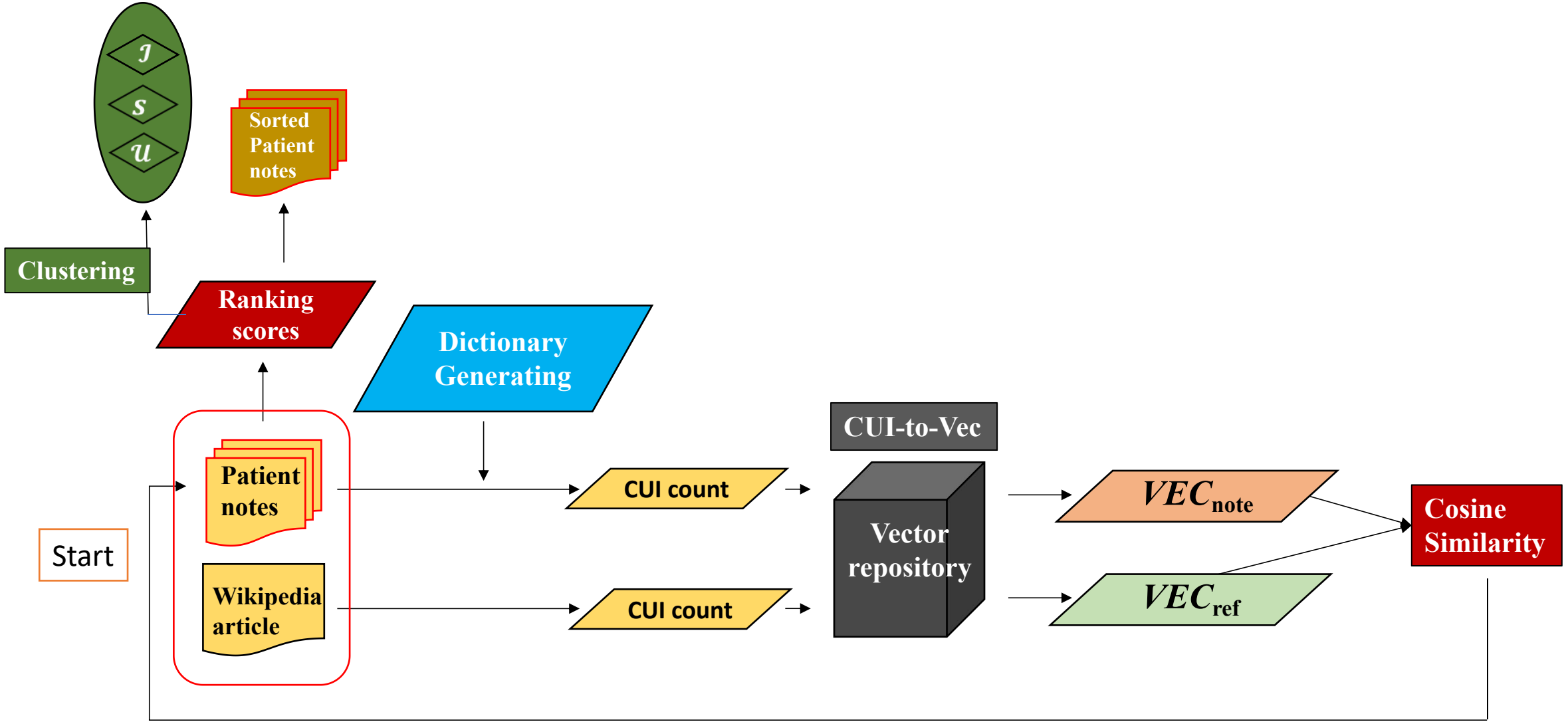


- **Obtain ranking scores** for notes

Note_ID	Ranking score
Note1	0.81
Note2	0. 44

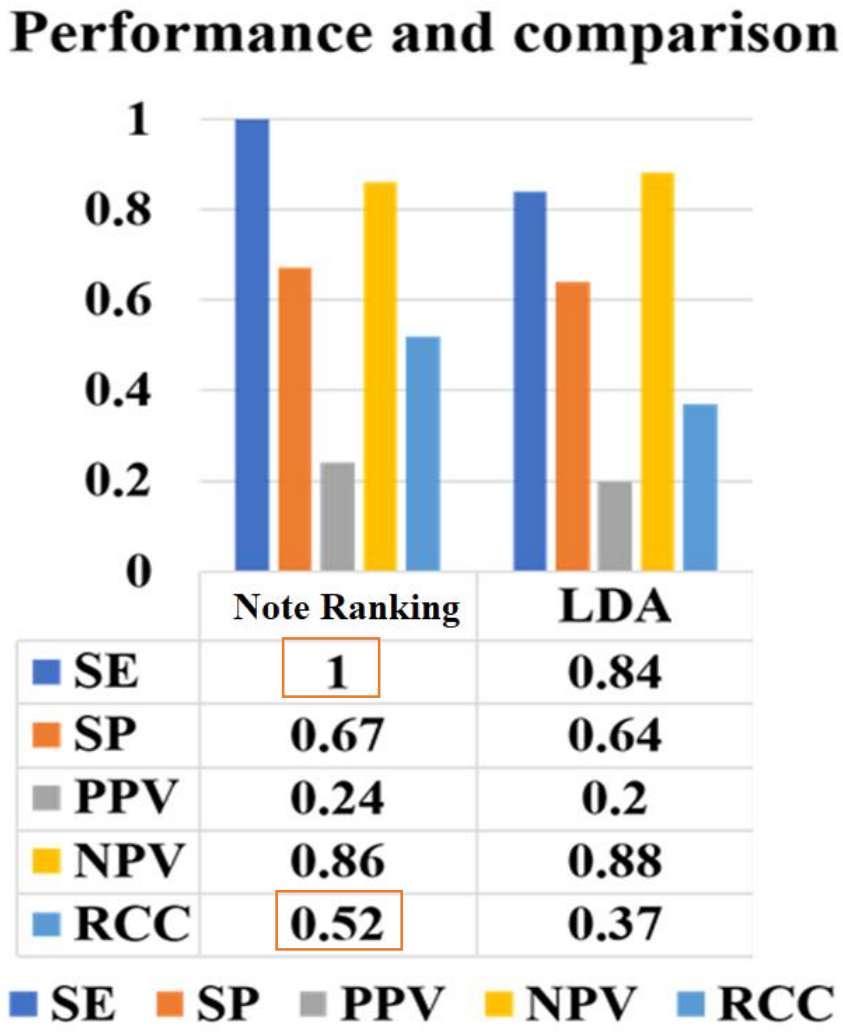
- **Cluster** ranking scores, **classify** the notes into uninformative (u), somewhat informative (S), and informative ($\mathcal{I}$)

Flow chart of note ranking algorithm



Evaluation and comparison

- **Gold standard**:
  - Manual review of 200 random notes



LDA: Latent Derichlet Allocation, SE: sensitivity, SP: Specificity, PPV: Positive Predictive Value, NPV: Negative Predictive Value, RCC: Rank Correlation Coefficient

Questions

Email: tcai1@bwh.Harvard.edu

THANK YOU