# Knowledge Node and Relation Detection

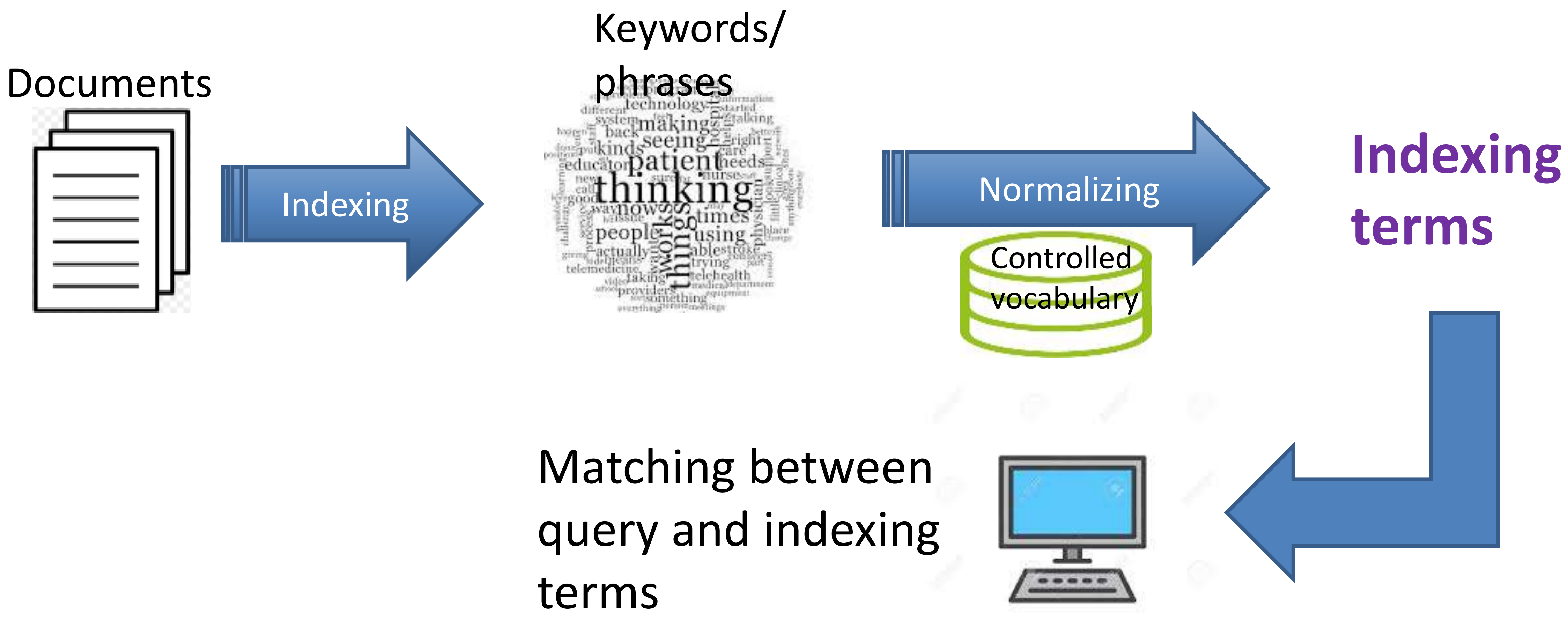
Jian Qin

School of Information Studies

Syracuse University

NKOS Workshop at DCMI/TPDL, Porto, Portugal, September 14, 2018

# It is all about
# subject content representation



Documents

Indexing

Keywords/ phrases

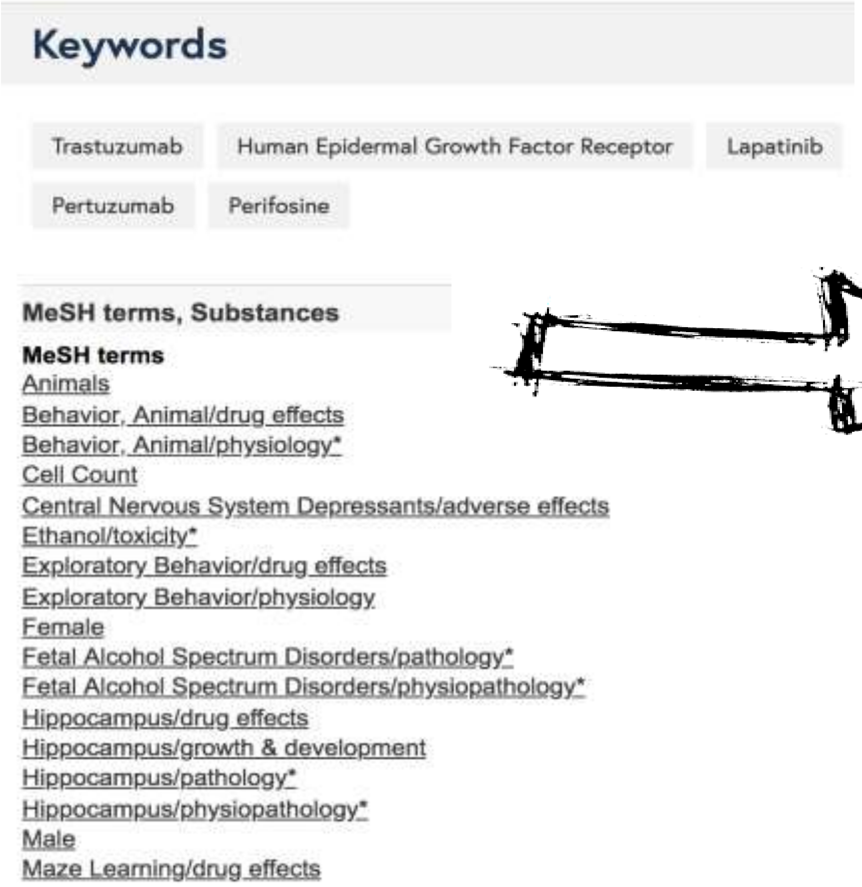
Normalizing

Controlled vocabulary

Indexing terms

Matching between query and indexing terms

# Increase interactivity by transforming the way knowledge is represented



## Keywords

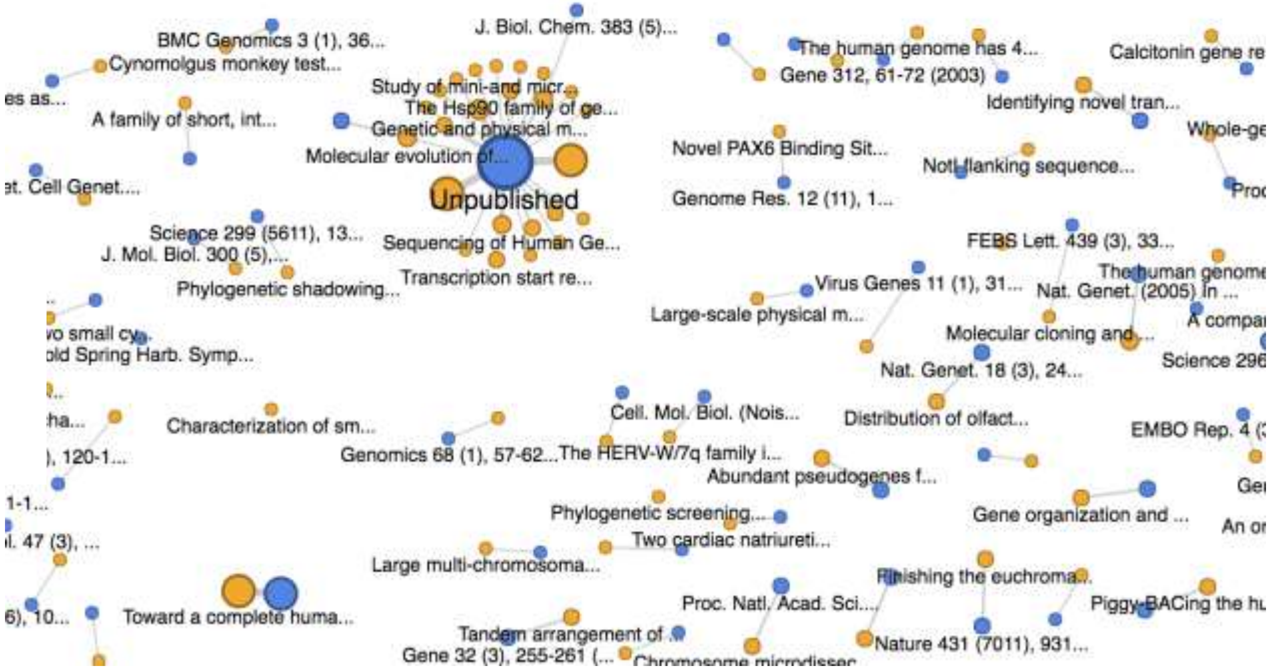
Trastuzumab | Human Epidermal Growth Factor Receptor | Lapatinib

Pertuzumab | Perifosine

### MeSH terms, Substances

**MeSH terms**

- Animals
- Behavior, Animal/drug effects
- Behavior, Animal/physiology*
- Cell Count
- Central Nervous System Depressants/adverse effects
- Ethanol/toxicity*
- Exploratory Behavior/drug effects
- Exploratory Behavior/physiology
- Female
- Fetal Alcohol Spectrum Disorders/pathology*
- Fetal Alcohol Spectrum Disorders/physiopathology*
- Hippocampus/drug effects
- Hippocampus/growth & development
- Hippocampus/pathology*
- Hippocampus/physiopathology*
- Male
- Maze Learning/drug effects

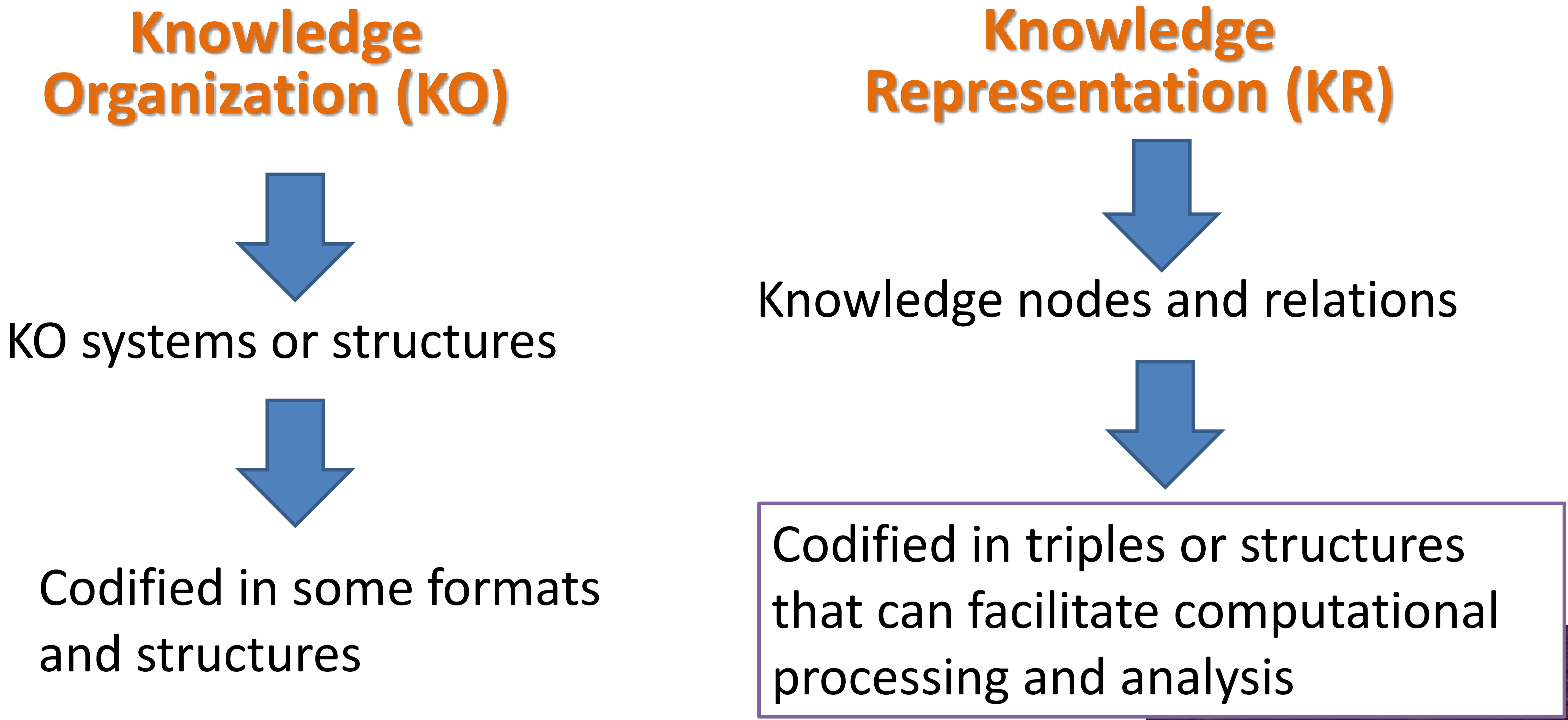


https://fusiontables.google.com/DataSource?docid=1Gs7wXxBl5TeiUrsV3MBoJNHouJEqjMk-ZSxmrsOC#chartnew:id=4

# Limitations

- Lack of rich relations between concepts (or other types of things) beyond scope relations
  - Between publications and data
  - Between datasets in different data repositories
  - Inside data and/or publications:
    - Between different types of entities
    - Between different topics
    - …
- Discrete terms from indexing process that must rely on relations defined in controlled vocabularies to show relations

# In linked data age…



## Knowledge Organization (KO)

KO systems or structures

Codified in some formats and structures

## Knowledge Representation (KR)

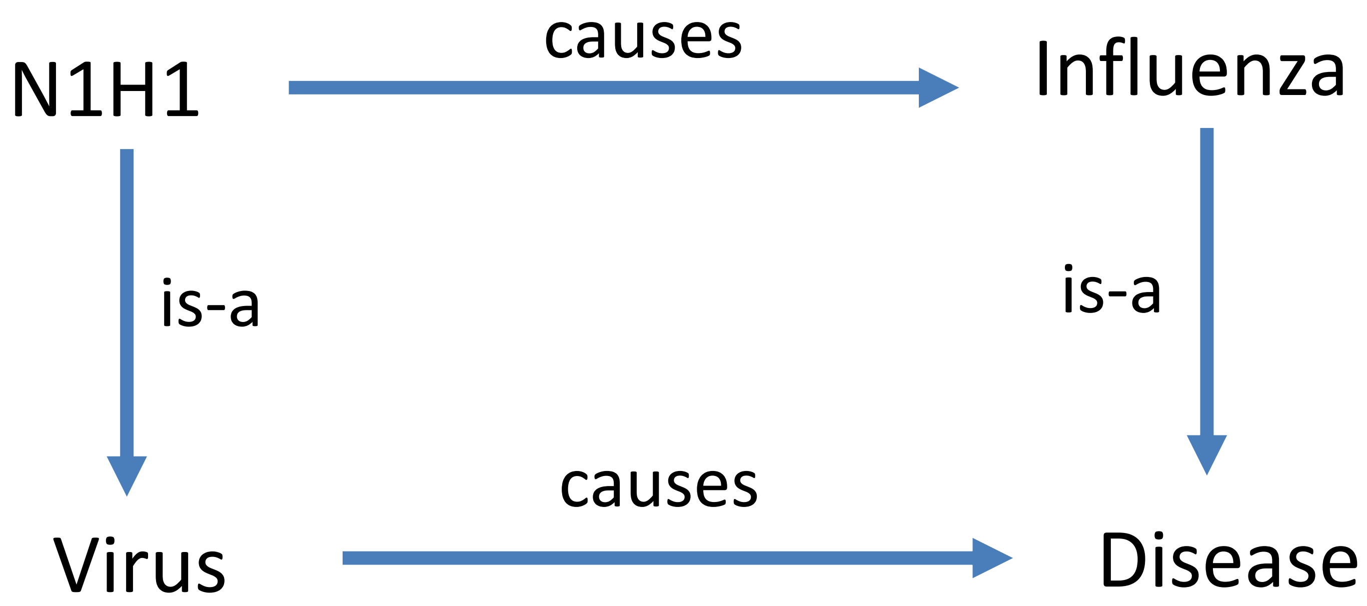
Knowledge nodes and relations

Codified in triples or structures that can facilitate computational processing and analysis

# How knowledge is represented?

- Currently two practices:
  - From natural language in full-text documents
    - Traditional indexing
    - Natural language processing and machine learning
  - From existing KOS through remodeling and restructuring
    - Converting existing KOS into linked data service (LCSH, MeSH, AAT)
    - Transform legacy data into linked data (e.g., library linked data)

# Paradigm shift from term-based representation to node-relation representation



Concept detection is relatively straightforward with help of KOS. Relation detection, however, has become the jewel in the crawn (or bottleneck problem) for representation knowledge in linked data age.

# Experiment

- 30 documents from PubMed
- Hand annotated 150 sentences to identify knowledge nodes (k-nodes) and relations in format [k-node(A), relation, k-node(B)]
- Indexing software used: MetMap and SemRep
  - Both support concept detection backed by UMLS
  - SemRep supports relation extraction
- Evaluation of results used Bilingual Evaluation Understudy (BLEU) and cosine similarity algorithm

# Research questions

- To what extent manually annotated and automatically generated k-nodes and relations are similar or dissimilar?

- What are some of the patterns of agreement and/or disagreement between the two sets of results?

- How can human intelligence (human-intervened k-node and relation recognition) be translated into machine intelligence for more accurate knowledge representation?

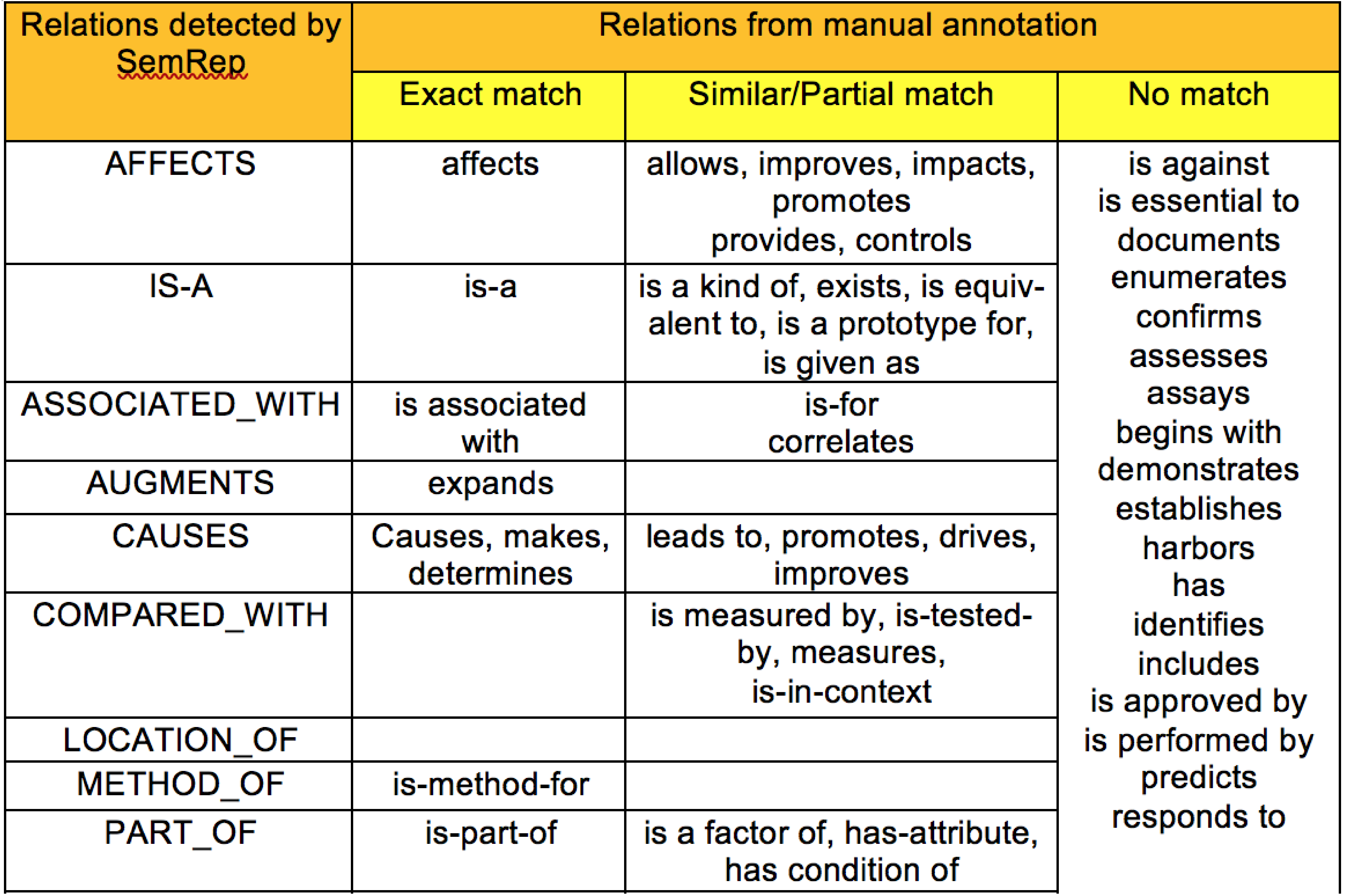
# Findings: degree of abstraction

| Sentence | Manually annotated k-nodes | MetaMap extracted k-nodes | SemRep extracted k-nodes |
|---|---|---|---|
| Unlike most pathologic testing, which serves as an adjunct to establishing a diagnosis, the results of HER2 testing stand alone in determining which patients are likely to respond to trastuzumab, a monoclonal antibody against HER2. | • pathologic testing<br>• HER2 testing<br>• monoclonal antibody<br>• trastuzumab<br>• HER2<br>• diagnosis | • pathologic testing<br>• results of her2 testing<br>• respond to trastuzumab<br>• results of her2 testing<br>• a monoclonal antibody against her2<br>• diagnosis | • pathologic<br>• testing<br>• HER2<br>• testing<br>• trastuzumab<br>• monoclonal antibody<br>• diagnosis |

# Findings: degree of abstraction

| Sentence | Manually annotated k-nodes | MetaMap extracted k-nodes | SemRep extracted k-nodes |
|---|---|---|---|
| At present, several preanalytic factors, including the time from tissue removal to tissue fixation, are underappreciated as important variables that have the potential to negatively impact the consistency and reliability of HER2 testing. | time from tissue removal to tissue fixation<br>preanalytic factor<br>HER2 testing<br>preanalytic factor<br>consistency<br>reliability | time from tissue removal<br>tissue fixation<br>several preanalytic factors<br>reliability of her2 testing<br>several preanalytic factors<br>consistency | time<br>removal<br>tissue fixation<br>factors<br>HER2<br>testing<br>consistency |

# Findings: degree of abstraction



| Relations detected by SemRep | Relations from manual annotation | | |
|---|---|---|---|
| | Exact match | Similar/Partial match | No match |
| AFFECTS | affects | allows, improves, impacts, promotes<br>provides, controls | is against<br>is essential to<br>documents<br>enumerates<br>confirms<br>assesses<br>assays<br>begins with<br>demonstrates<br>establishes<br>harbors<br>has<br>identifies<br>includes<br>is approved by<br>is performed by<br>predicts<br>responds to |
| IS-A | is-a | is a kind of, exists, is equivalent to, is a prototype for, is given as | |
| ASSOCIATED_WITH | is associated with | is-for<br>correlates | |
| AUGMENTS | expands | | |
| CAUSES | Causes, makes, determines | leads to, promotes, drives, improves | |
| COMPARED_WITH | | is measured by, is-tested-by, measures,<br>is-in-context | |
| LOCATION_OF | | | |
| METHOD_OF | is-method-for | | |
| PART_OF | is-part-of | is a factor of, has-attribute, has condition of | |

# Findings: types of k-nodes and relations

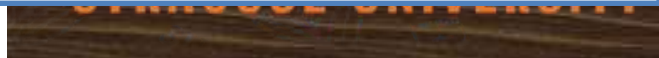
*Simple k-node relations*: two simple k-nodes are connected by a direct relation in the form of a single verb: $A \rightarrow B$

(amplification_of_HER2_gene, promotes, receptor_activation)
(tumor, harbors, HER2_molecular_alteration)

*Compound k-node relations*: refers to situation where one k-node is related to more than one k-node that has the same or different relations: $A \rightarrow (B_1 \ldots B_n)$

(overexpression_of_receptor, mediates, biology_behavior_of_HER2-positive_tumor_cells)
(overexpression_of_receptor, mediates, clinical_behavior_HER2-positive_tumor_cells)
(overexpression_of_receptor, drives, proliferation_of_tumor_cells)
(overexpression_of_receptor, drives, survival_of_tumor_cells)

(overexpression_of_receptor, mediates, (biology_behavior_of_HER2-positive_tumor_cells, clinical_behavior_HER2-positive_tumor_cells))
(overexpression_of_receptor, drives, (proliferation_of_tumor_cells, survival_of_tumor_cells))



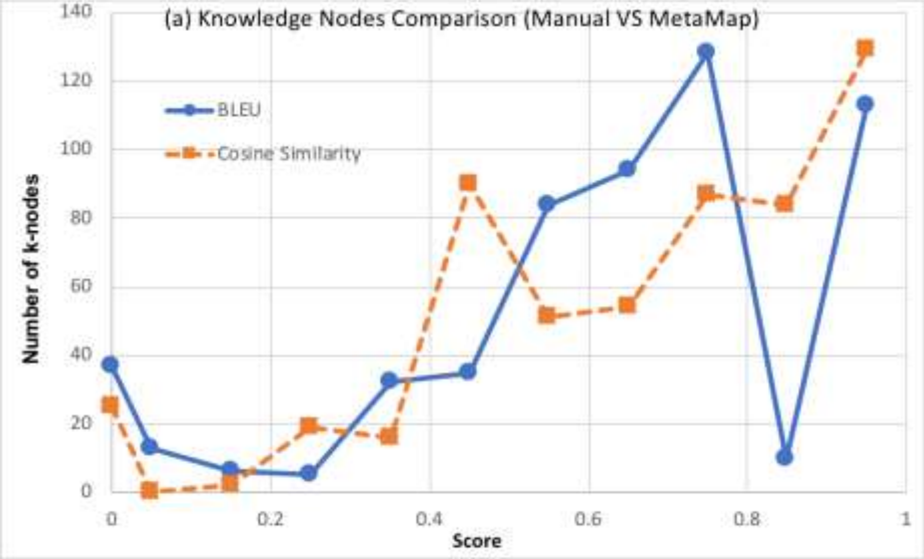
# Findings: types of k-nodes and relations (cont'd)

*Complex k-node relations*: multiple k-nodes and the relations chained together by "bridge" k-nodes: A→(B→C)

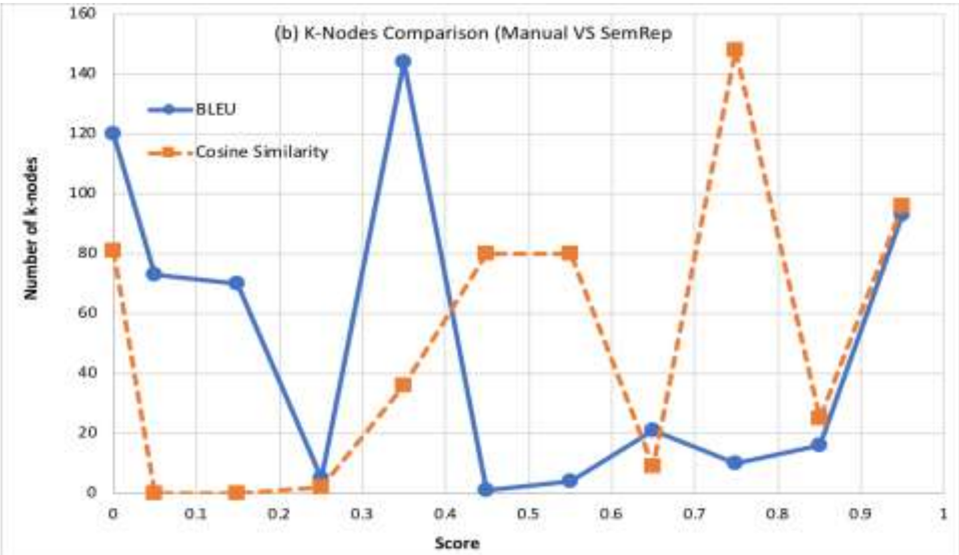
(HER2_testing, determines, (patient, responds-to, trastuzumab))

(trastuzumab, is-a, monoclonal antibody against HER2)

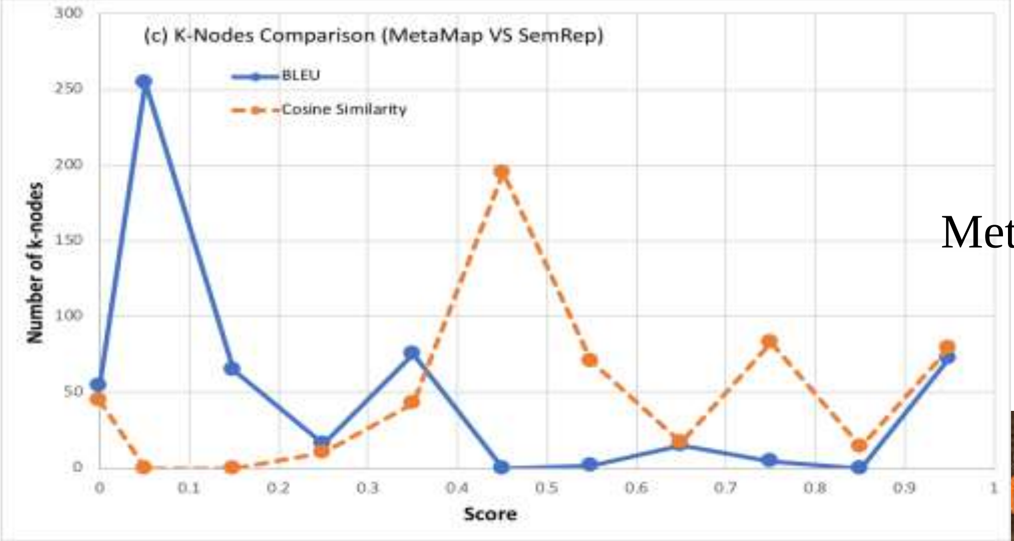
# Evaluation scores for three k-node detection methods



Manual vs. MetaMap

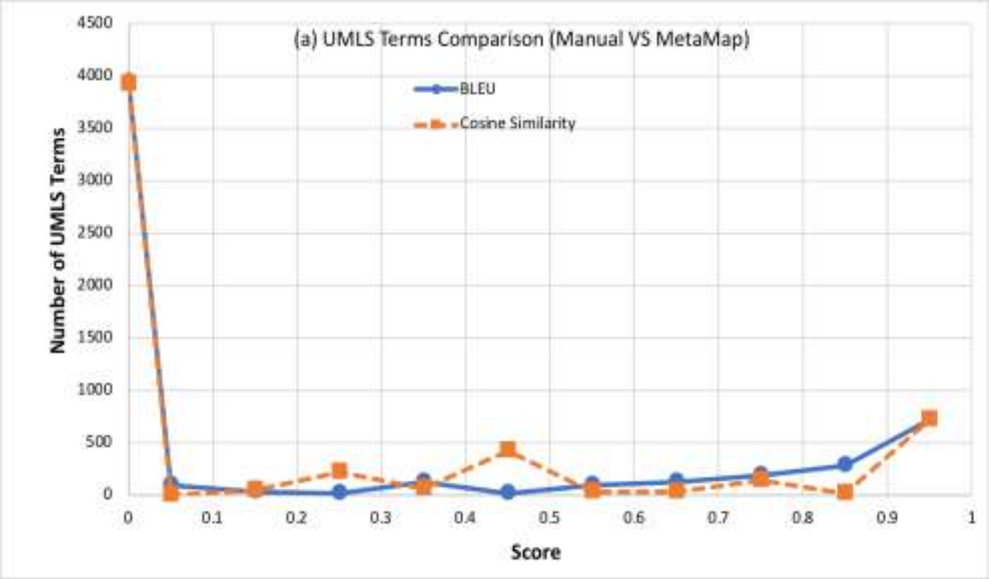


Manual vs. SemRep

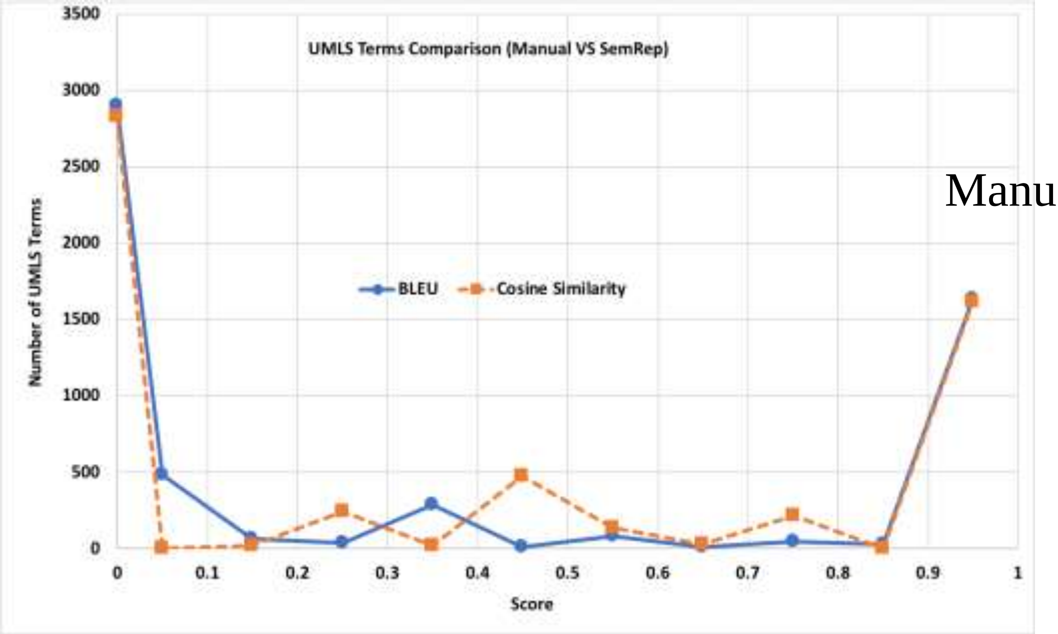


MetaMap vs. SemRep

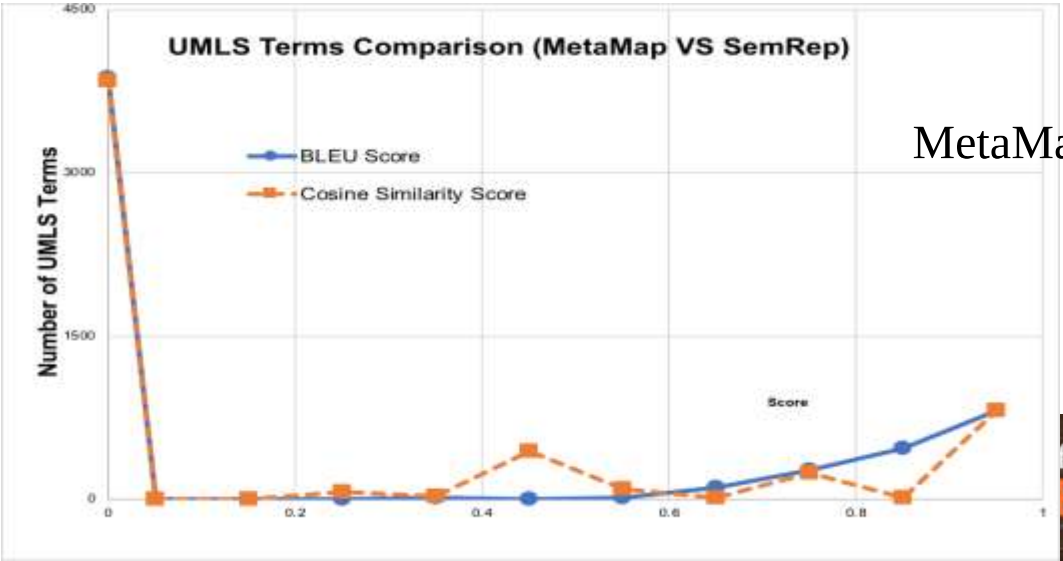
# Evaluation scores for UMLS term matching



Manual vs. MetaMap



Manual vs. SemRep



MetaMap vs. SemRep

# Discussion and conclusion

- knowledge node and relation recognition from full-text documents is highly challenging, yet critically important in the big data era.
- Each of k-node and relation detection methods has different areas of strengths and limitations.
- Automatic tools have a long way to go
- K-node and relation representation facilitates knowledge network generation

Thank You